



BT4015 Geospatial Analytics Final Group Project

(Group 10)

<u>Member Name</u>	<u>Matric No</u>
Ching Zheng Ing	A0202003I
Leung Hoi Kit Alvin	A0201468J
Tan Tian Le	A0188014H

Content Page

1. Details of Context of Data	2
2. Motivation	2
3. Problems or Insights that are going to be Solved or Discovered	4
4. Data Description	6
4.1 Data Pre-Processing	8
5. Exploratory Spatial Data Analysis (ESDA)	10
5.1 Data Visualisation	10
5.2 Density Plots	12
5.3 Point Pattern Analysis	15
6. Spatial Data Analysis	20
6.1 Buffers for Amenities	20
6.2 Hypothesis Testing	28
6.2.1 Average Nearest Neighbour (ANN)	30
6.2.2 Poisson Process Model	34
6.3 Spatial Autocorrelation	38
6.4 Geographically Weighted Regression (GWR)	39
6.4.1 Predicting Resale Prices	39
6.4.2 Predicting Number of HDBs in Area	41
7. Discussion and Conclusion	43
7.1 Discussion	43
7.2 Limitations and Further Improvements	43
7.2.1 Lack of Individual HDB Prices	43
7.2.2 Limited Amenities	44
7.2.3 Island Temperature	45
7.3 Conclusion	46
References	47

1. Details of Context of Data

In a small island of Singapore, it is common to see many high-rise residences, known as HDB flats, which are managed by the Housing and Development Board. They are scattered in the different regions and planning areas in Singapore, but are they created equal, or are there some planning areas where it is more conducive to live in? With this question in mind, the data that we have used will contain both spatial information and non-spatial information that could help us derive meaningful insights.

2. Motivation

HDB flats are what most people in Singapore would call a home, with around 80% of residents currently living in it (Hirschmann, 2021), as reflected in Figure 1 below.

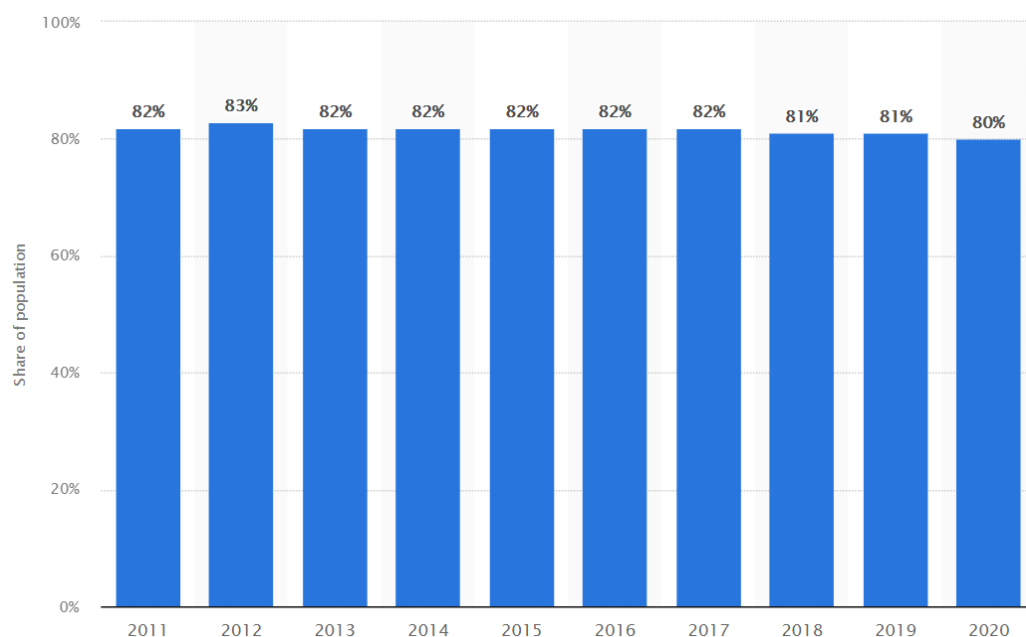


Figure 1: % of Residents in Singapore living in HDB flats from 2011 - 2020

This is a significant proportion of people in Singapore, which is expected, considering how small Singapore is. When we consider the possible areas that one can live in, the size of the liveable area shrinks even further. As of 2020, Singapore has a population size of 5.86 million, and this number is expected to increase, judging from the trend in the past 70 years, as reflected in Figure 2 below. This judgement was also made as the Singapore Government is planning ahead for population growth in the future. While it is not expected for our population to reach between 6.5 and 6.9 million by 2030, as indicated by the Population

White Paper in 2013, we can most definitely expect the population to continue growing steadily. This in turn leads to a higher demand for HDB flats for more residents to live in. Figure 3 shows the number of residents that are living in HDBs for the past 20 years, which was released by the Singapore Department of Statistics (DOS) in 2021, and we can see that the increasing trend coincides with Singapore's population in Figure 2 (Worldometer, n.d.).

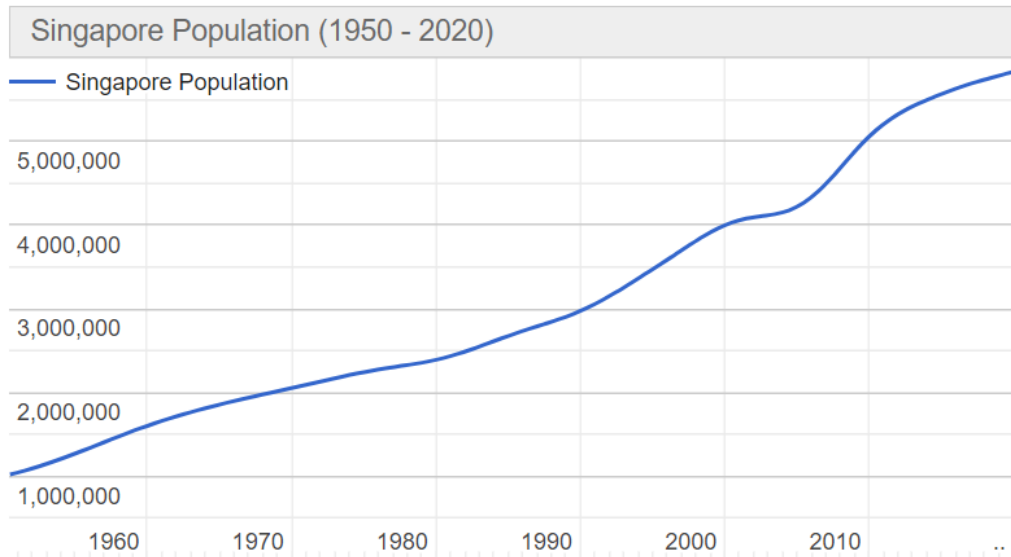


Figure 2: Singapore Population from 1950 to 2020

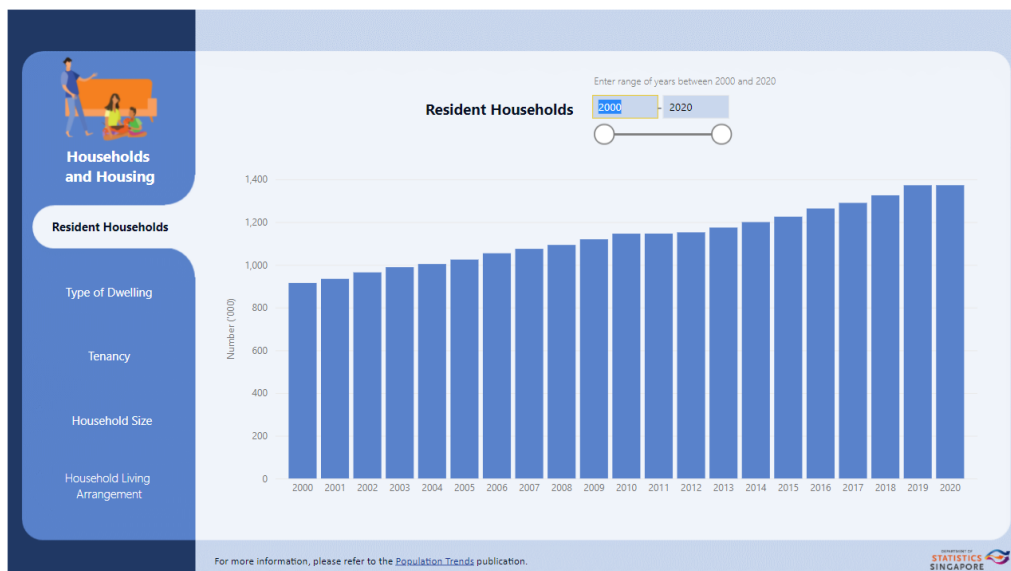


Figure 3: Number of Residents in Singapore living in HDB from 2000 to 2020

This trend is expected as high rise building is the best way to fit increasing population in a small island. In addition, we can expect this number to increase as more newly married couples are balloting for their future homes through the Built to Order (BTO) system, introduced in 2001. This also arises due to the changing culture where more married couples are no longer living with their parents and are now owning their own properties.

Recently, the waiting times for these slots have been lengthening, with BTO projects launching last year having an average waiting time of about four to five years. Considering that couples will spend a long time, or perhaps even a lifetime, in this home, it is extremely important for us to understand what would make a flat ideal for them. Consequently, we would like to investigate exactly what factors adults in Singapore would consider to be important when buying a home in Singapore.

3. Problems or Insights that are going to be Solved or Discovered

In this project, our group aims to explore how we can identify if a location is good and what makes a location good. To better help us with the analysis, we have made some critical assumptions. Firstly, we have assumed that the locations of existing HDBs are good, as these locations have been carefully selected and planned out by the Government. Despite this, there are certainly locations which are better than another to live at, which can be observed by the resale prices. Therefore, the next assumption that needs to be made is that price will be an indicator of how good a location is. Considering the law of demand and supply, the price of the HDBs at particular locations will increase if there is a high demand for it. For a location to have a high demand, it should be a location of what people would identify as good. This is a reasonable assumption as people are willing to pay more for what they want. Lastly, we will assume that amenities within the vicinity will be an indicator of a good location. Specifically, we will make the following hypothesis, as reflected in Figure 4 below.

Amenities	Hypothesis	Reasons
Bus Stops	HDBs located near bus stops are better locations.	Having bus-stops nearby will allow residents to easily travel to other locations in Singapore.
MRT and LRT Stations	HDBs located near MRT and LRT stations are better locations.	Having MRT and LRT stations nearby will allow residents to easily travel to other locations in Singapore.
Pre-Schools	HDBs located near pre-schools are better locations.	Having pre-schools nearby will allow parents to easily send their children to pre-schools, or to allow the children to safely travel to and fro themselves.
Primary and Secondary Schools	HDBs near primary and secondary schools are better locations.	Having primary and secondary schools nearby will allow parents to apply for priority admissions for better schools, and also allow their children to easily travel to and fro of school.
Market and Hawker	HDBs near market and hawker centres are better locations.	Having market and hawker centres nearby will allow residents to easily purchase their meals, and to buy groceries home. This is especially important if they are carrying heavy groceries.
Shopping Malls	HDBs near shopping malls are better locations.	Having shopping malls nearby will allow residents to conveniently shop and relax in an air-conditioned environment. This is especially so with the popularity of retail therapy in a highly stressed environment, such as Singapore.

Figure 4: Hypothesis made based on Amenities

To summarise the above tables, the closer the HDBs are to various amenities, the better they are as a location. By determining if these amenities are indeed indicators of what makes a location optimal for HDBs, we are aiming to derive some meaningful insights, such as the best locations purchase HDBs based on the nearby amenities, as well as the best locations to purchase them based on both the nearby amenities and the resale value. This is critical for people to consider as it is an expensive purchase, and there is often a budget for such purchases. Therefore, they will want to get the best location with their limited budgets. In addition, this should give some insights to the Government on what kind of amenities are important to the population, and what they can build to improve the existing HDBs.

4. Data Description

There are different types of data that have been used in this project, which includes shapefiles, Keyhole Markup Language (KML) files and the typically used CSV files. These files capture a great variety of information, as reflected in Figure 5 below. While Singapore can be divided into regions, planning areas and sub-zones, we have eventually chosen the planning areas to be the base of our analysis. This is due to the regions being over generic, with only 5 regions in the whole of Singapore. It would have led to an unmeaningful analysis that may not have given any critical insights that would benefit the Government and the residents. On the other hand, sub-zones are too detailed, with over 300 sub-zones in Singapore. There are in fact many sub-zones which are unheard of, such as Pang Sua and Saujana, and including all of these details would not necessarily provide valuable information to residents, when they do not even know the locations. In addition, some of the other data sources that we have worked with are categorised by planning areas, such as the estimated population, and it would affect our analysis if we were to use sub-zones instead. Hence, the planning area of Singapore was selected to form the base of our analysis.

S/N	Data Description	File Type	Attributes Used	Source
1	Planning Areas of Singapore	SHP File (Polygons)	PLN_AREA_N, SHAPE_AREA, geometry	Data.gov.sg
2	HDB Locations/Market & Hawker Locations	CSV	Block number, Street, Market & hawker tag, Town	Data.gov.sg (HDB Property Information)
3	Bus Stop Locations	SHP File (Points)	geometry	LTA Data Mall - Bus Stop Location
4	MRT and LRT Exits	KML File (Polygons)	geometry	Data.gov.sg
5	Pre-school Locations	KML File (Points)	geometry	Data.gov.sg
6	School Locations	CSV	Address and Postal Code	Data.gov.sg (General Information of Schools)
7	Shopping Mall Locations	N/A	List of Shopping Malls in Singapore (Central, East, North, North East, North West, South, West)	Wikipedia
8	HDB Resale Prices	N/A	Average price of 4 room flats in the different planning area	https://www.propertyguru.com.sg/property-guides/singapore-hdb-resale-flat-price

				es-28448
9	Weather (Temperature)	JSON	Station latitude and longitude, weather readings over time	Data.gov.sg (Realtime Weather Readings)
10	Estimated Population by Location	CSV	Town or Estate, HDB Resident Population (Number of Persons)	Data.gov.sg (Estimated Resident Population in HDB Flats, by Town)

Figure 5: Data used for Analysis

4.1 Data Pre-Processing

Due to the nature of the data, there is a lot of pre-processing that needs to be done before we can start doing any spatial analysis on them.

For many of the datasets, such as the Primary and Secondary schools and HDB locations, we only had a list of Addresses or postal codes while for others such as the shopping mall locations we only had a list of names and general areas.

In order to plot these features on a Singapore map it was necessary to process the data and geocode each element with an associated latitude and longitude so that we would know where they were located. Failure to do so would prevent us from conducting the majority of the spatial analysis and so it was necessary to find a way to do so.

Luckily, we found a solution with OneMap, a geocoding API provided for free by the Singapore Land Authority. Using the tools provided by the API allowed us to label all our data points in python with their latitudes and longitudes and export the resultant shapefile which we could then directly import into R.

```
In [12]: def get_xy(address):
address_name = address.replace(" ", "+")
r = requests.get("https://developers.onemap.sg/commonapi/search?searchVal=" +
address +
"&returnGeom=Y&getAddrDetails=N&pageNum=1")
elevations = r.json()
df = pd.json_normalize(elevations['results'])
##df_location.head()
try:
#return elevations
ll = list(df[['LATITUDE', 'LONGITUDE']].iloc)
return float(ll[0][0]), float(ll[0][1])
except:
return "Error"
```

Figure 6: Function used to Geocode Locations using the OneMap API

Out[15]:	school_name	postal_code	Combined Address	latlong	lat	long
0	ADMIRALTY PRIMARY SCHOOL	738907	11 WOODLANDS CIRCLE 738907	(1.4426347903311, 103.800040119743)	1.442635	103.80004
1	ADMIRALTY SECONDARY SCHOOL	737916	31 WOODLANDS CRESCENT 737916	(1.44589068910993, 103.802396196596)	1.445891	103.802396
2	AHMAD IBRAHIM PRIMARY SCHOOL	768843	10 YISHUN STREET 11 768843	(1.43315271543517, 103.832942401086)	1.433153	103.832942
3	AHMAD IBRAHIM SECONDARY SCHOOL	768928	751 YISHUN AVENUE 7 768928	(1.43605975368804, 103.829718690077)	1.43606	103.829719
4	AI TONG SCHOOL	579646	100 Bright Hill Drive 579646	(1.3605834338904, 103.833020333988)	1.360583	103.83302
...
341	ZHANGDE PRIMARY SCHOOL	169485	51 Jalan Membina 169485	(1.28421153855474, 103.825951884637)	1.284212	103.825952
342	ZHENGHUA PRIMARY SCHOOL	679002	9 Fajar Road 679002	(1.37942673617052, 103.76970317201)	1.379429	103.769703
343	ZHENGHUA SECONDARY SCHOOL	677741	91 SENJA ROAD 677741	(1.38836583415352, 103.765510638527)	1.388366	103.765511
344	ZHONGHUA PRIMARY SCHOOL	556095	12 SERANGOON AVENUE 4 556095	(1.36026072476019, 103.869712517383)	1.360261	103.869713
345	ZHONGHUA SECONDARY SCHOOL	556123	13 SERANGOON AVENUE 3 556123	(1.34840720846208, 103.869430481414)	1.348407	103.86943

346 rows x 6 columns

```
In [23]: school = [Point(xy) for xy in zip( school_df['long'], school_df['lat'])]
crs = {'init': 'epsg:4326'}
sg_df = gpd.GeoDataFrame(school_df, crs = crs, geometry = school)

In [27]: singapore_map = gpd.read_file('Planning Area Census/Planning_Area_Census2010.shp')
singapore_map = singapore_map.to_crs(epsg=4326)
fig,ax = plt.subplots(figsize = (15,15))
sg_df.plot(ax = ax, markersize = 20, color = "blue", marker = "o", label = "School")
singapore_map.plot(ax = ax, alpha = 0.3, color = "grey")
```

Figure 7: Geocoded School locations from Address and Postal Code Information

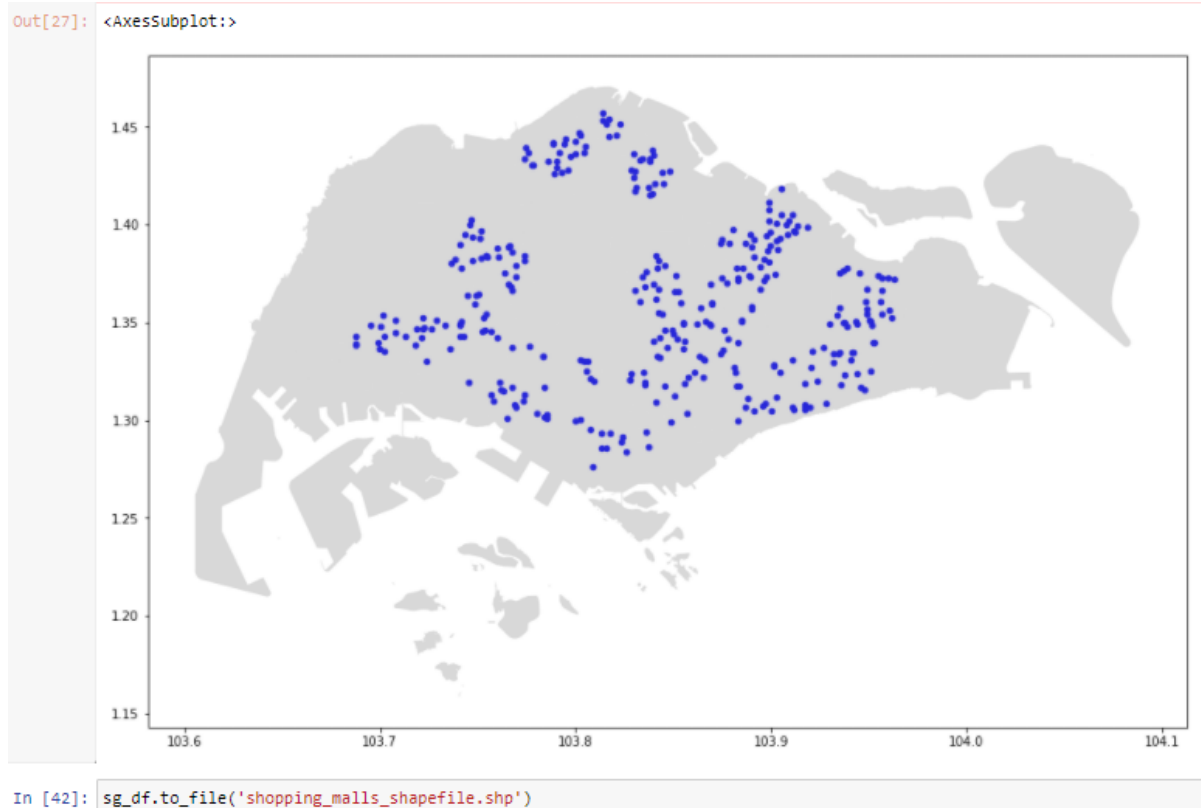


Figure 8: Checked Coordinates of Schools by plotting on Singapore Map and exporting geoDataframe as Shapefile

5. Exploratory Spatial Data Analysis (ESDA)

5.1 Data Visualisation

We chose to study Singapore according to the planning areas decided by the Urban Redevelopment Authority. Since the implementation of these boundaries, other government ministries and departments, including the housing development board have also increasingly adopted these boundaries for their administrative purposes, which is why we have chosen to use this as our base map in hopes that the use of the planning areas will help add context to our findings.

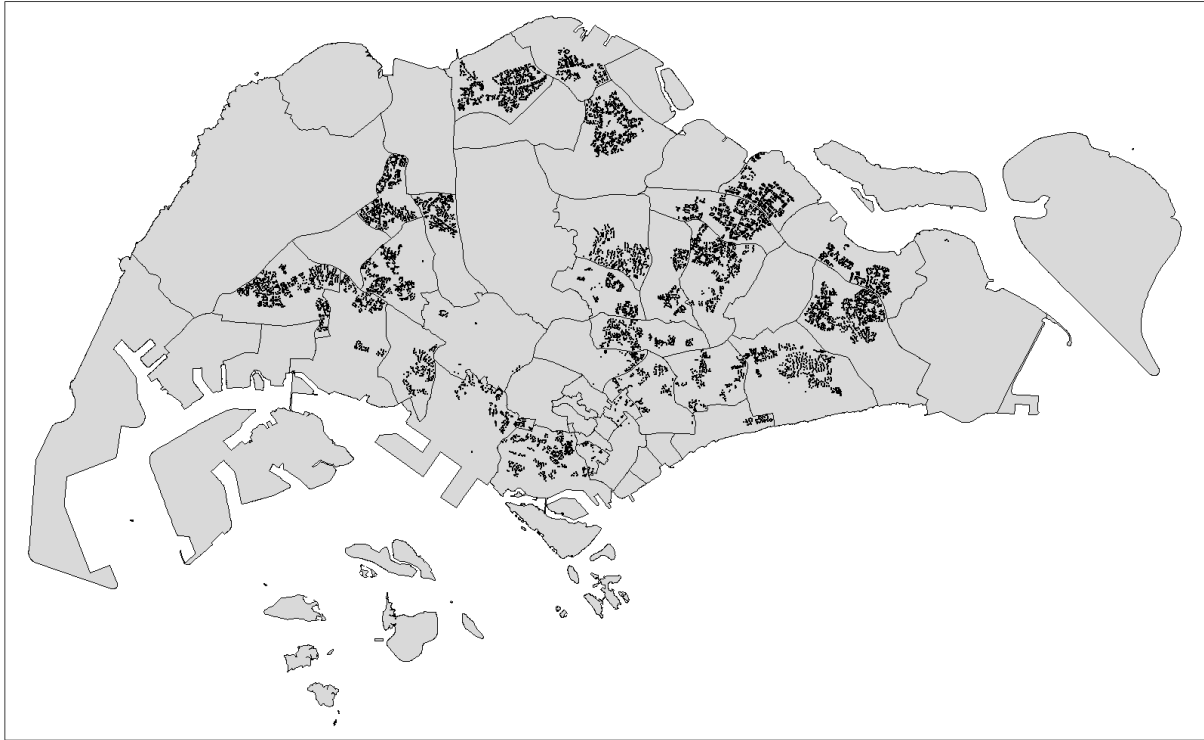


Figure 9: Plot of HDB Blocks within Singapore Planning Areas

In addition to the HDB locations, we also used a dataset of HDB resale prices for different areas of Singapore, since we wanted to be as up to date as possible while keeping standards consistent we have chosen to only use the resale value of 4 room HDB flats from the 2nd Quarter of 2021.

Because of this, and the fact that not all planning areas contain HDBs, there are some locations where there is no data available for resale price. Luckily, the vast majority of HDBs fall within zones that have price labels, and hence we can still use it as an indicator of how good a HDB location is.

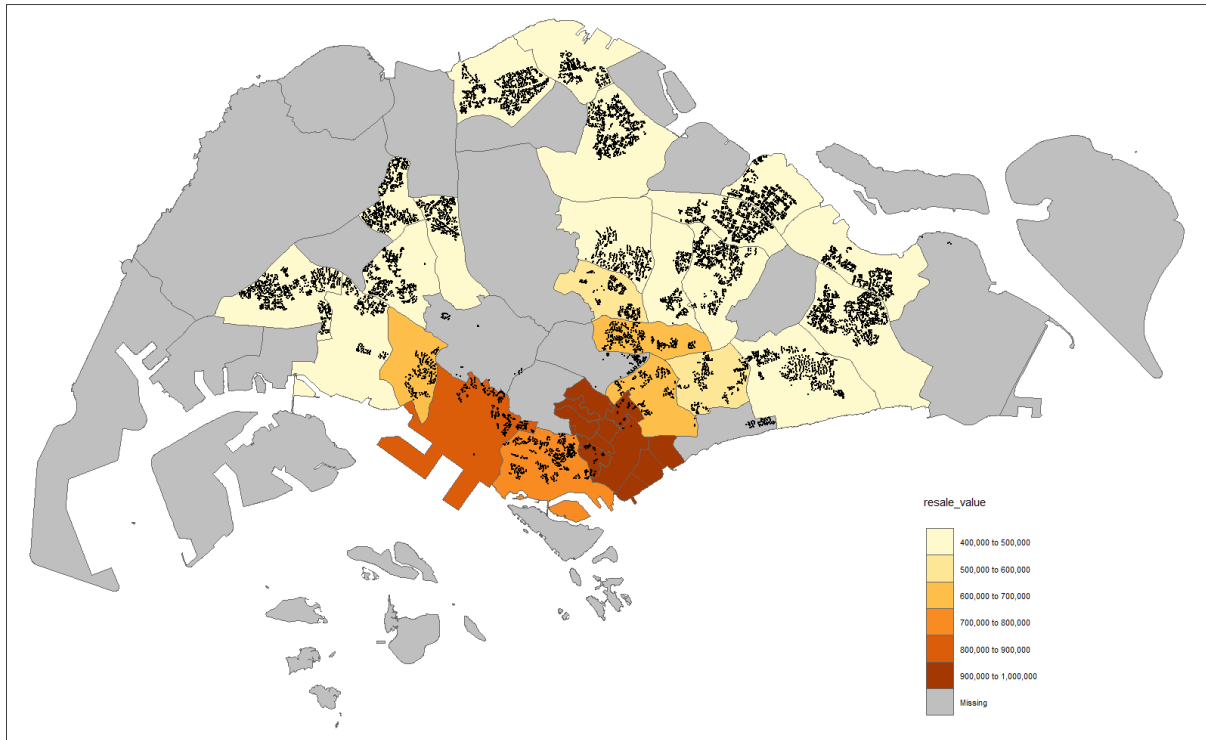


Figure 10: Plot of HDB blocks within Singapore Planning Areas, with Planning Areas distinguished by Colour indicating Area Resale Value

Moreover, to get a sense of where the different amenities we decided to study are located within Singapore, we also plotted out where their individual locations are within each of the planning areas.

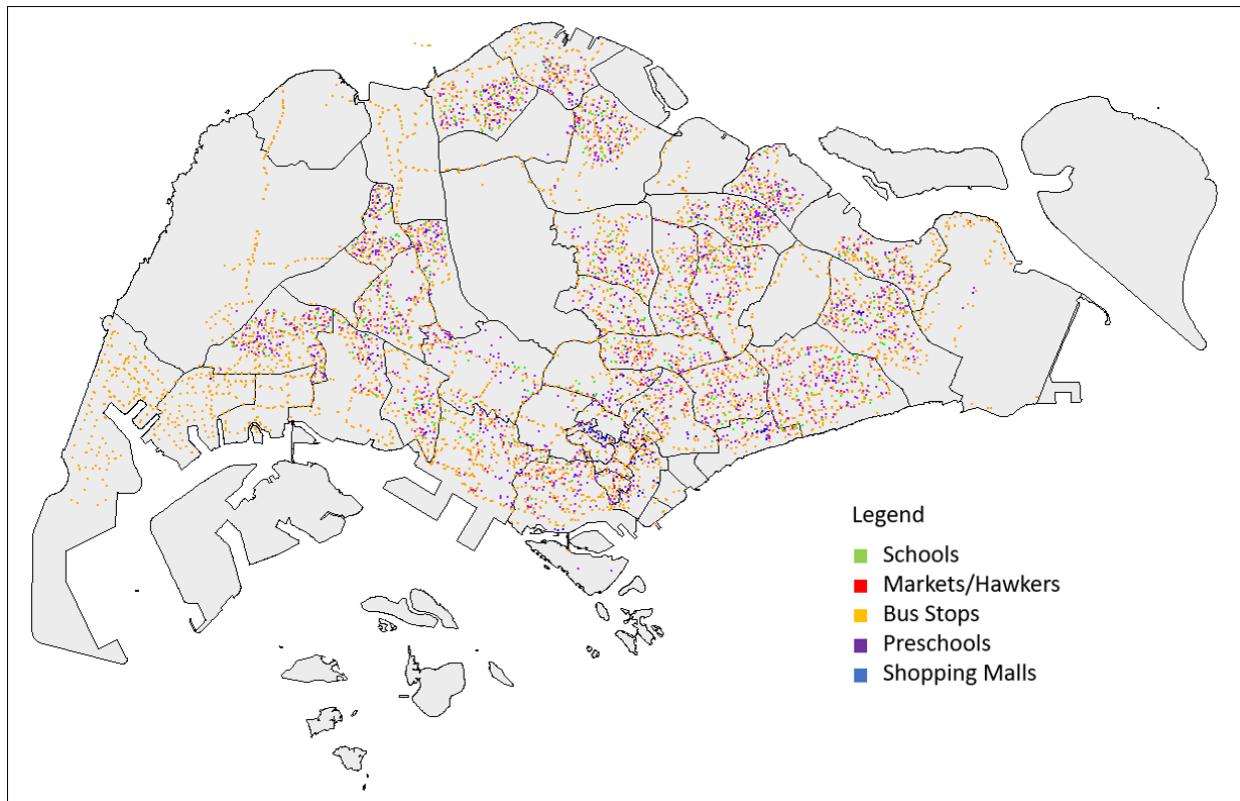


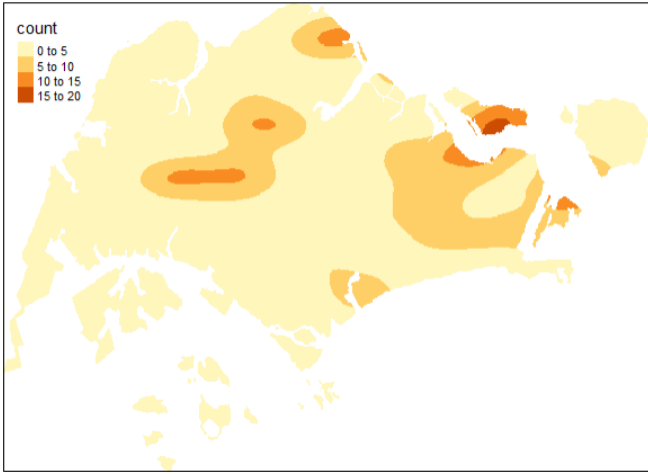
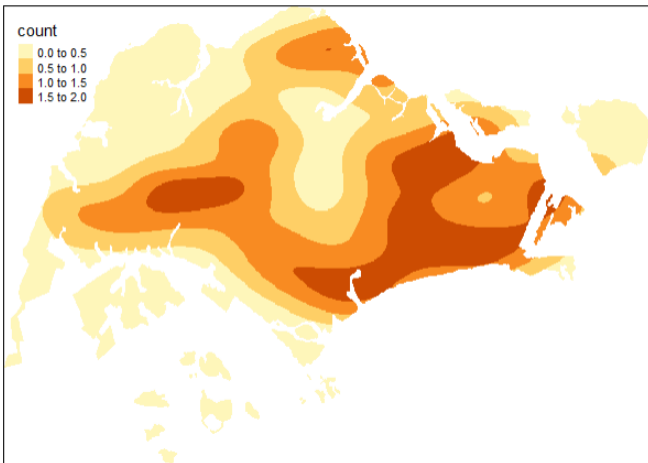
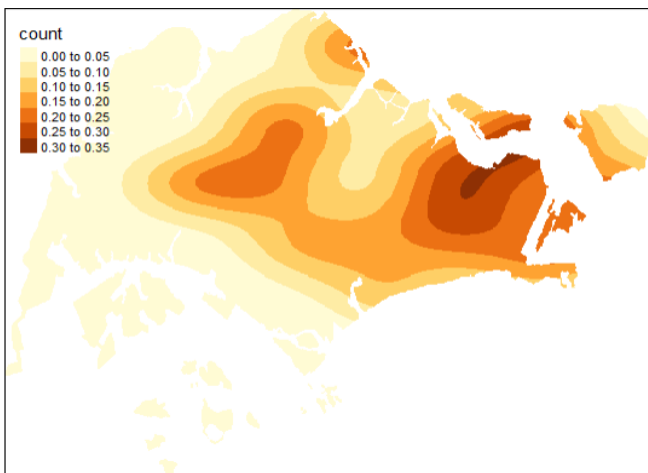
Figure 11: Plot of the Different Amenities and where they are located within the different Planning Areas

Now that we had an idea of where everything is within Singapore, we decided to investigate which areas of Singapore had higher densities of each amenity for the next part of our exploratory data analysis.

5.2 Density Plots

One of our assumptions was that clustered areas, or areas with higher densities of amenities would make better locations for building and buying a HDB, this is logical because we can assume that homeowners would naturally like to live closer to an amenity rather than further away, and hence living in an area with a high density of an amenity would mean having it more easily accessible.

In the table below we have used the Smooth Map function to get the density of different amenities in order to observe where exactly are the areas in Singapore with high density, as well as to see which amenity density is also consistent with HDB density.

Feature	Plot	Description
HDB	 <p>count</p> <ul style="list-style-type: none"> 0 to 5 5 to 10 10 to 15 15 to 20 	There are 3 main areas of high density, the Eastern region, Northern region and Central area, with there also being a slightly more dense area in the South.
Bus Stops	 <p>count</p> <ul style="list-style-type: none"> 0.0 to 0.5 0.5 to 1.0 1.0 to 1.5 1.5 to 2.0 	The areas of high density are similar in the Eastern region, Northern region and Central area. However, the area that these regions cover is much larger and even overlaps at their boundaries.
Schools	 <p>count</p> <ul style="list-style-type: none"> 0.00 to 0.05 0.05 to 0.10 0.10 to 0.15 0.15 to 0.20 0.20 to 0.25 0.25 to 0.30 0.30 to 0.35 	The areas of high density are similarly in the Eastern region, Northern region and Central area. However, the area that these regions cover is much larger and even overlaps at their boundaries.

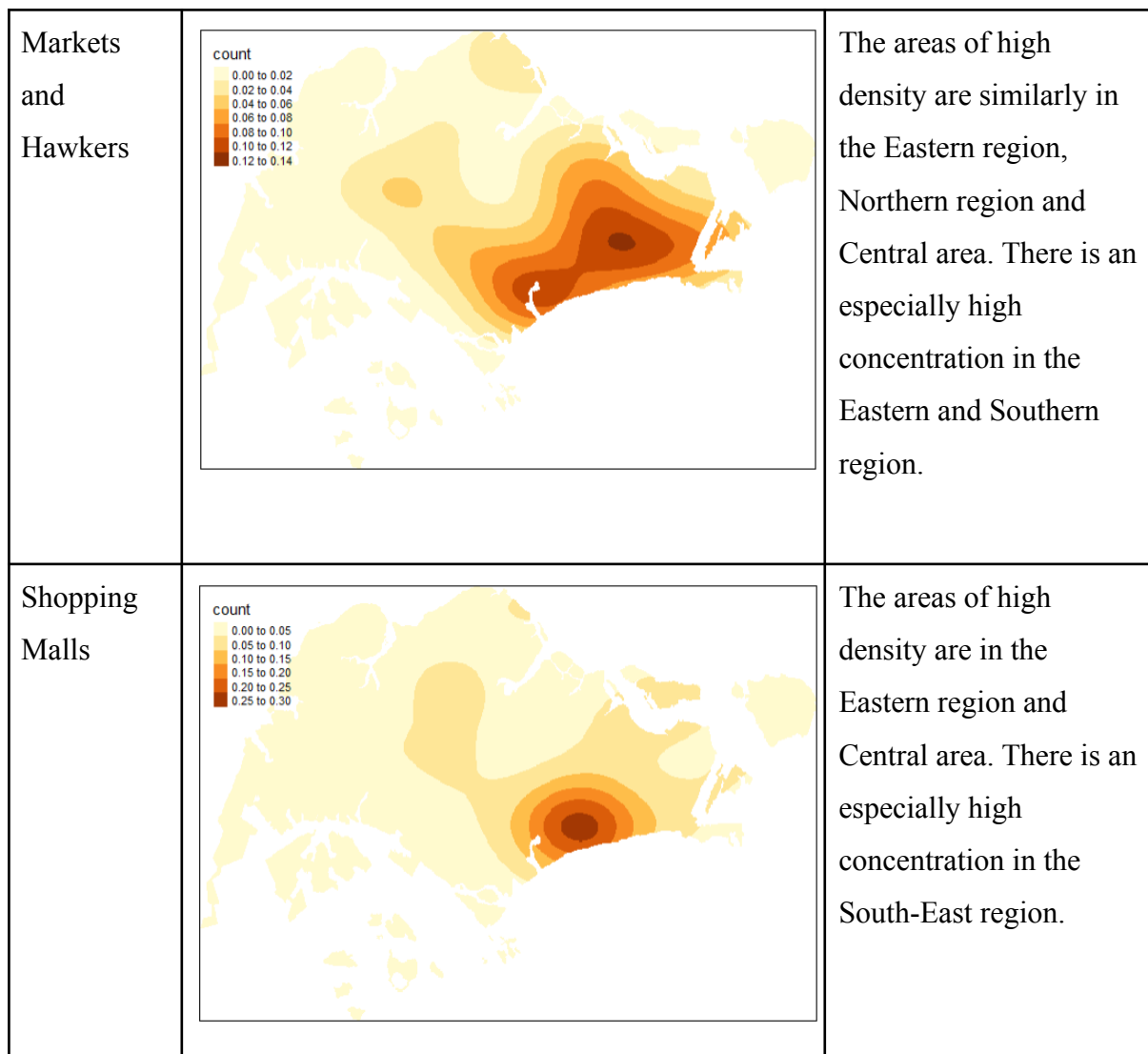


Figure 12: Density Plots and Observations

Overall from observing the different density plots it can be seen that generally all our amenities have areas of higher density in the East, North and Central areas, which aligns with the density of HDBs. We can conjecture that these areas are the main designated residential areas of Singapore which is why all the plots have these areas in common.

Additionally, what we noted for some specific amenities are that for bus stops, while there are areas of high density these areas are much larger and overlap between regions, which indicates the importance that Singapore placed on having bus stops available throughout the island so as to ensure our transport network has a wide coverage. This effect is also observed for schools, albeit to a smaller extent which makes sense given the benefit that students gain from having a school that is easily accessible regardless of where one stays.

The “coverage” effect we observed here, relating to having an amenity within close distance will be explored further later in our analysis.

5.3 Point Pattern Analysis

Next, in order to confirm if there truly is clustering within the features we are studying, we have to confirm that they are not situated similar to an independent random process (IRP), also called complete spatial randomness (CSR).

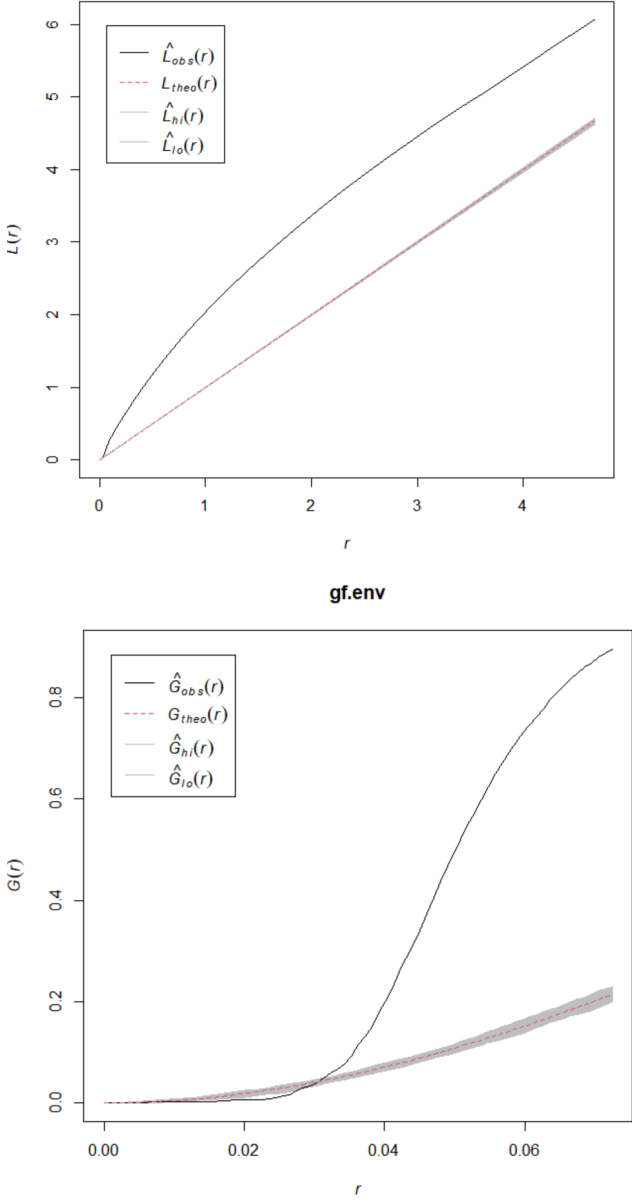
In order to check if that is the case we can use the K function, which summarizes the distance between points for all distances. However, since one problem with the K function is that the shape of the function tends to curve upward making it difficult to see small differences between K and K_{expected} , we used the L function to address this issue.

A shortcoming of the K function (and by extension the L function) is its cumulative nature which makes it difficult to know at exactly which distances a point pattern may stray from K_{expected} since all points up to distance r can contribute to $K(r)$.

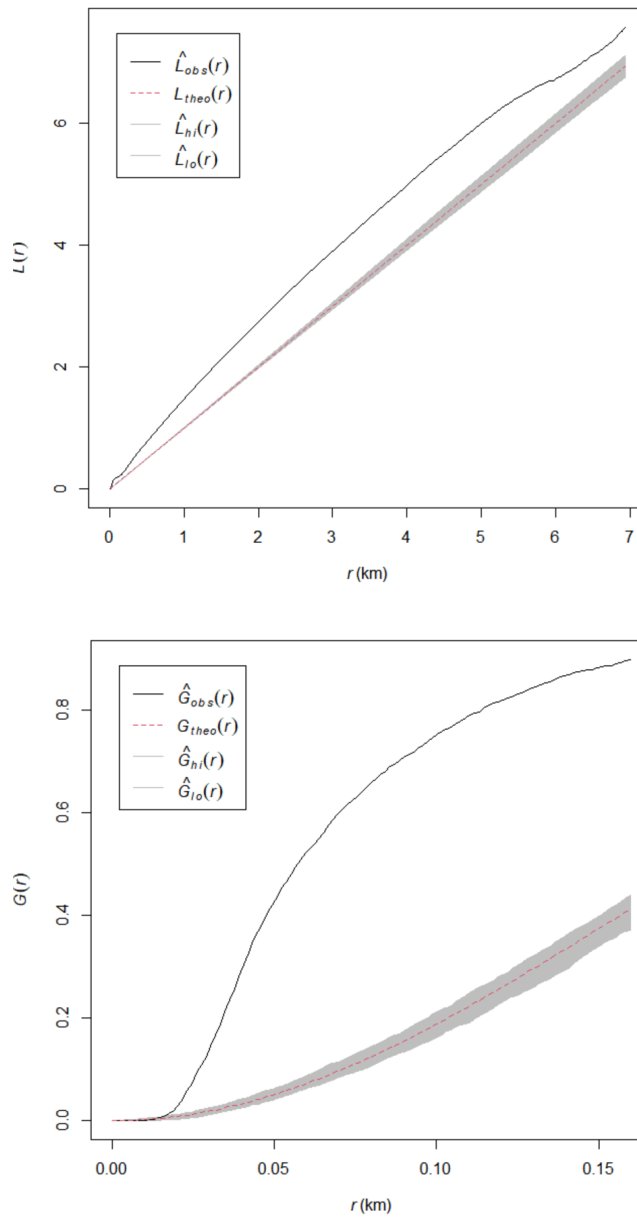
The pair correlation function, g , is a modified version of the K function where instead of summing all points within a distance r , points falling within a narrow distance band are summed instead.

Consequently, we decided to plot the L and G function plots for all of our features to see if they are consistent with CSR or not.

- If $g(r) = 1$, then the inter-point distances (at and around distance r) are consistent with CSR.
- If $g(r) > 1$, then the points are more clustered than expected under CSR.
- If $g(r) < 1$, then the points are more dispersed than expected under CSR.

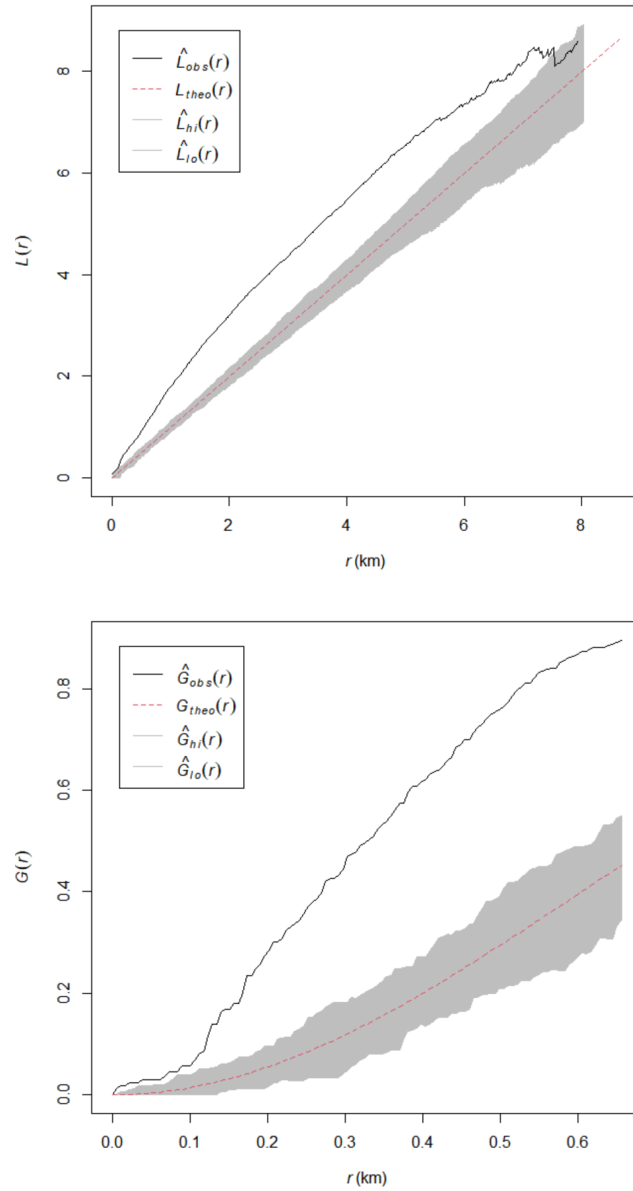
Feature	L and G Function Plots	Description
HDB	 <p>The figure consists of two vertically stacked plots. The top plot shows the L-function $L(r)$ on the y-axis (ranging from 0 to 6) against r on the x-axis (ranging from 0 to 4). It contains four curves: $\hat{L}_{obs}(r)$ (solid black), $L_{theo}(r)$ (dashed red), $\hat{L}_{hi}(r)$ (solid grey), and $\hat{L}_{lo}(r)$ (solid light grey). The observed curve is significantly higher than the theoretical curve. The bottom plot shows the G-function $G(r)$ on the y-axis (ranging from 0.0 to 0.8) against r on the x-axis (ranging from 0.00 to 0.06). It contains four curves: $\hat{G}_{obs}(r)$ (solid black), $G_{theo}(r)$ (dashed red), $\hat{G}_{hi}(r)$ (solid grey), and $\hat{G}_{lo}(r)$ (solid light grey). The observed curve is significantly higher than the theoretical curve.</p>	<p>For both the L and G functions it can be seen that the $L(r) > L(r)_{expected}$ and $G(r) > G(r)_{expected}$</p>

Bus Stops



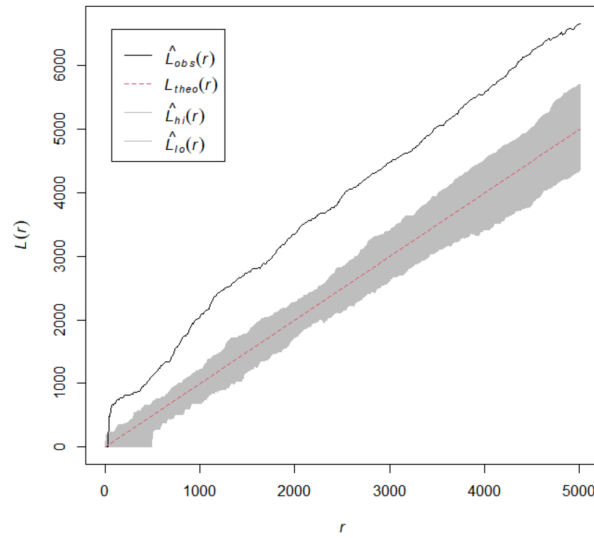
For both the L and G functions it can be seen that the $L(r) > L(r)_{\text{expected}}$ and $G(r) > G(r)_{\text{expected}}$

Schools

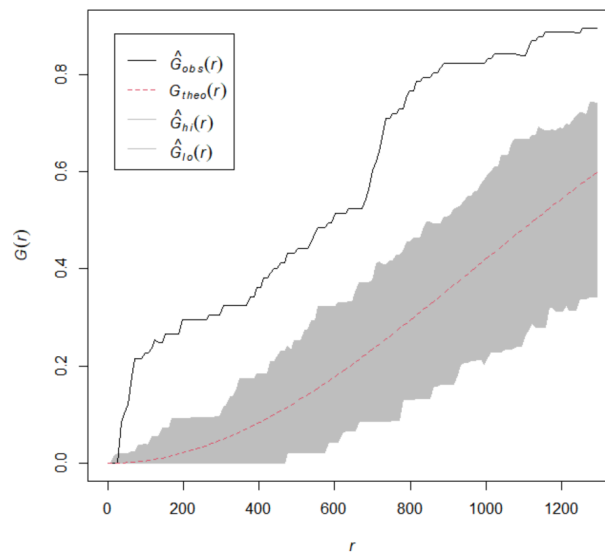


For both the L and G functions it can be seen that the $L(r) > L(r)_{\text{expected}}$ and $G(r) > G(r)_{\text{expected}}$

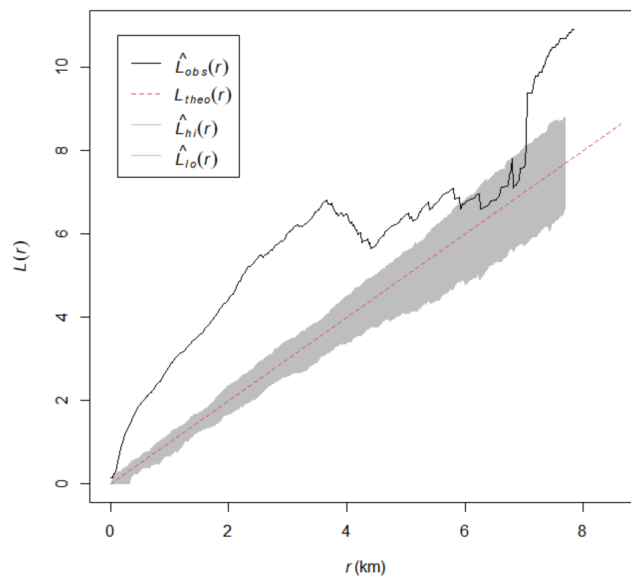
Market
and
Hawker



For both the L and G functions it can be seen that the $L(r) > L(r)_{\text{expected}}$ and $G(r) > G(r)_{\text{expected}}$



Shopping
Mall



For both the L and G functions it can be seen that the $L(r) > L(r)_{\text{expected}}$ and $G(r) > G(r)_{\text{expected}}$

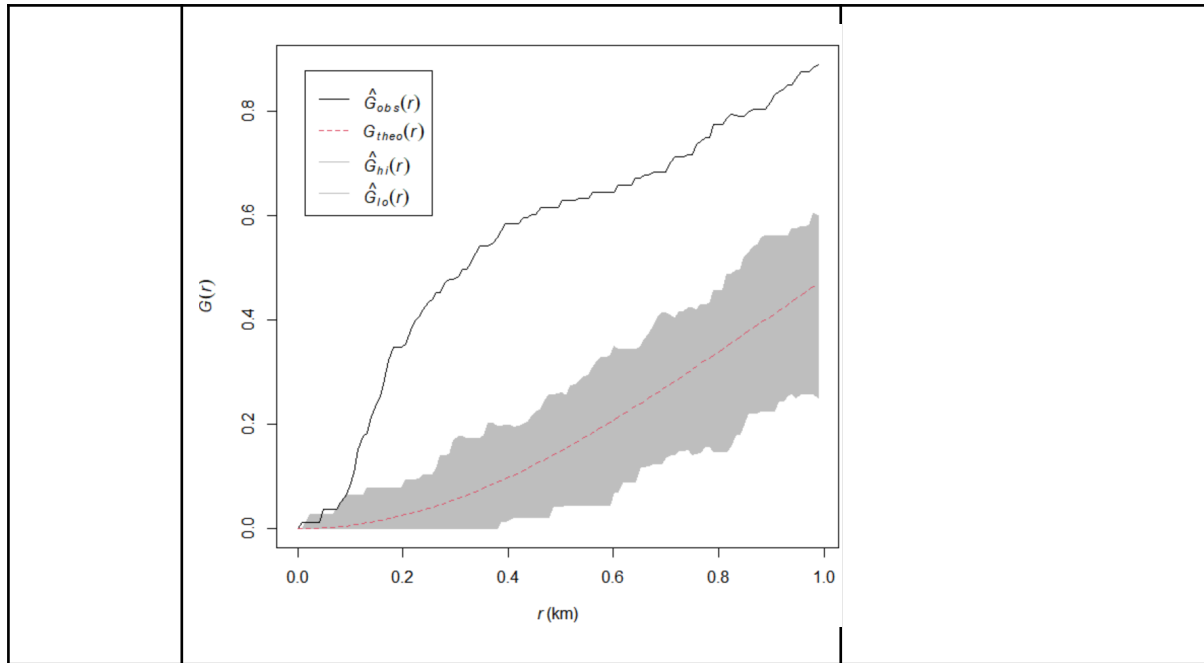


Figure 13: Table of L and G Function Plots

Overall we can conclude that all of our features are more clustered than expected under CSR, which means that the areas they are clustered at should be regarded as better locations than others.

6. Spatial Data Analysis

6.1 Buffers for Amenities

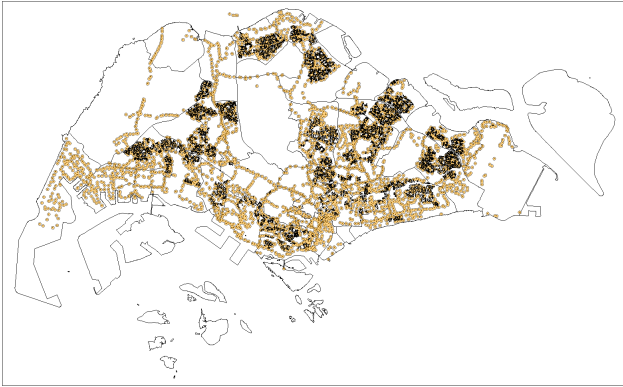
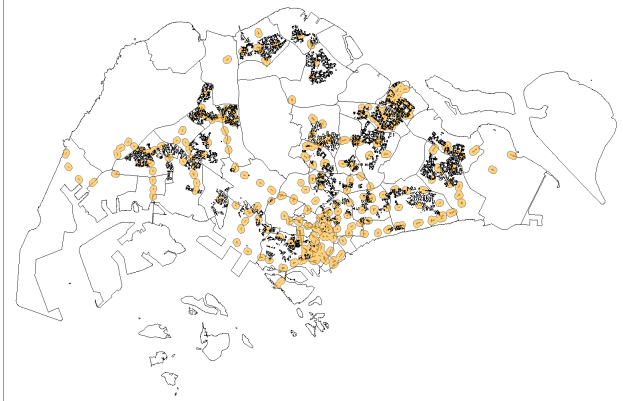
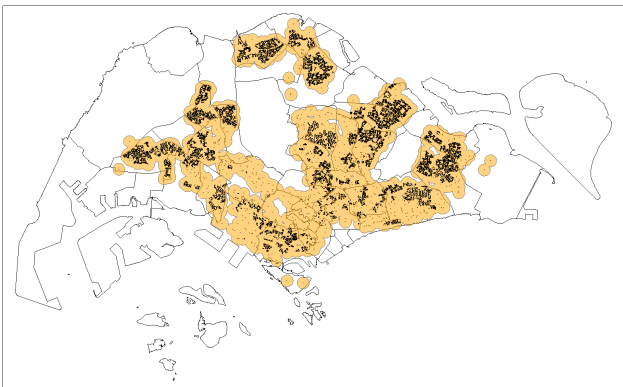
We created buffer zones for the amenities mentioned above. The distance of the buffer varied according to each amenity.

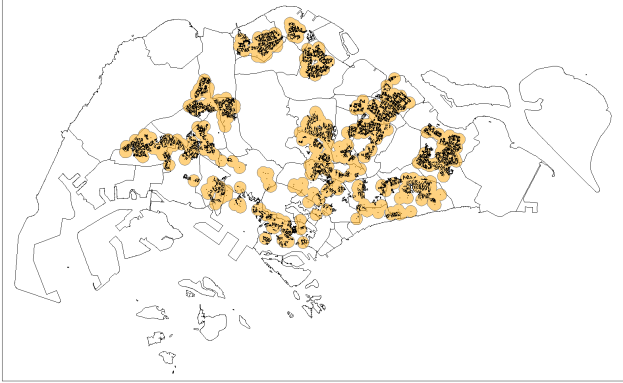
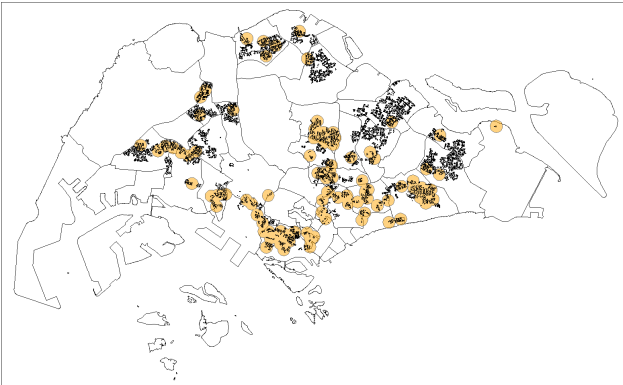
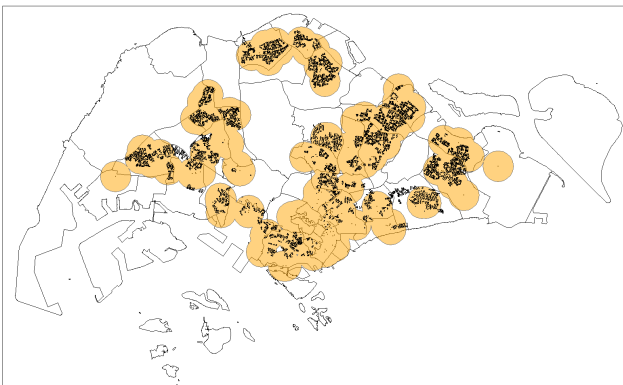
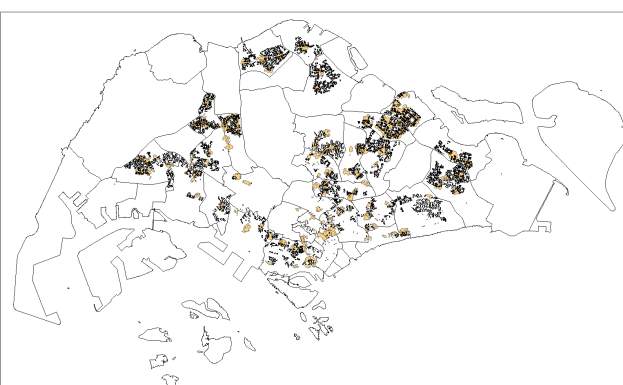
Amenities	Buffer Distance	Reasons
Bus Stops	500m	Ideally bus stops should be walking distance from HDB buildings. According to Statista (2021), walking distance in Singapore is roughly 500m.
MRT and LRT Stations	1km	As nearby busses are likely to bring commuters to train stations, a higher leniency is given for both MRT and LRT Stations.

Pre-Schools	2km	<p>Residents within 1 - 2km of primary schools are given priority for admission (MOE, 2021). The same buffer distance is used for primary, secondary and pre school as both amenities have relatively similar purposes.</p> <p>The same distance is used for market and hawker. This is because schools, markets and hawkers are amenities that are visited very frequently, almost daily. Hence we assume a similar distance would be considered ideal for HDB.</p>
Primary and Secondary Schools		
Market and Hawker		
Shopping Malls	5km	<p>As compared to schools, markets and hawkers, shopping malls have a buffer distance as the frequency of visiting a shopping mall would generally be lower.</p>

Figure 14: Reasons for Buffer Distance

The number of HDB that fall within these buffers were then calculated, alongside their percentages. The buffer zones were also intersected to create a combined buffer zone. It is worthy to note that close to 100% of HDBs fall within 2km of preschools, while 89% of them fall within 2km of primary and secondary school. This likely illustrates the importance of education among residents of Singapore.

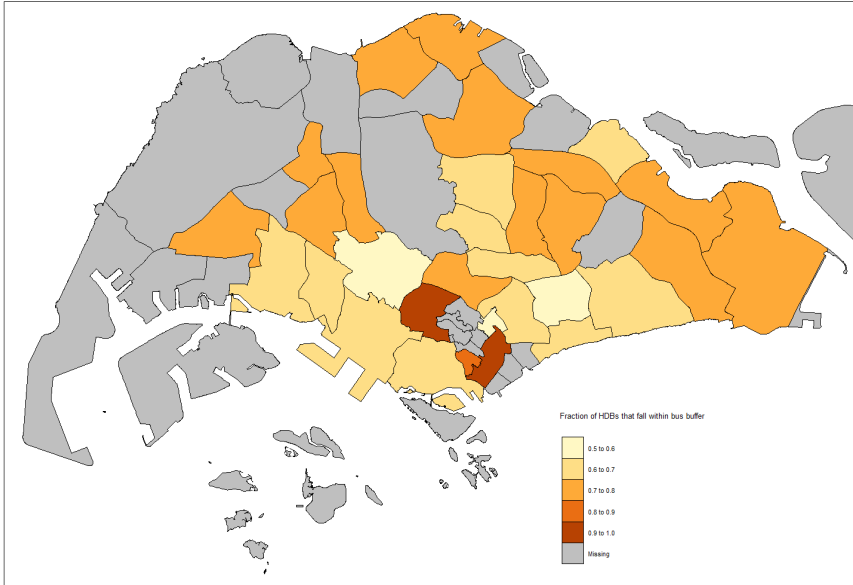
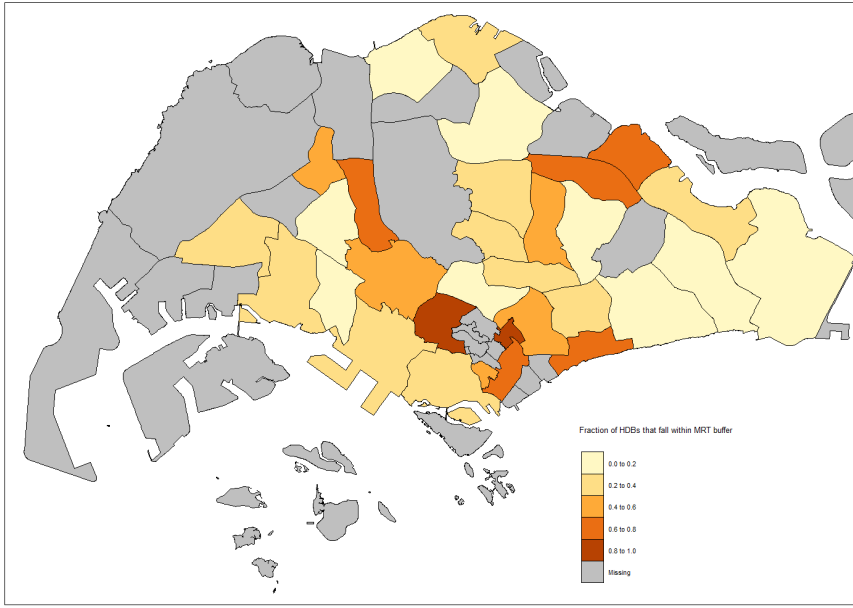
Amenities	Plot	No. of HDB that falls within the Buffer Zone	Percentage of HDBs that fall within Buffer Zone
Bus Stops		7220	71%
MRT and LRT Stations		3210	32%
Pre-Schools		10149	100%

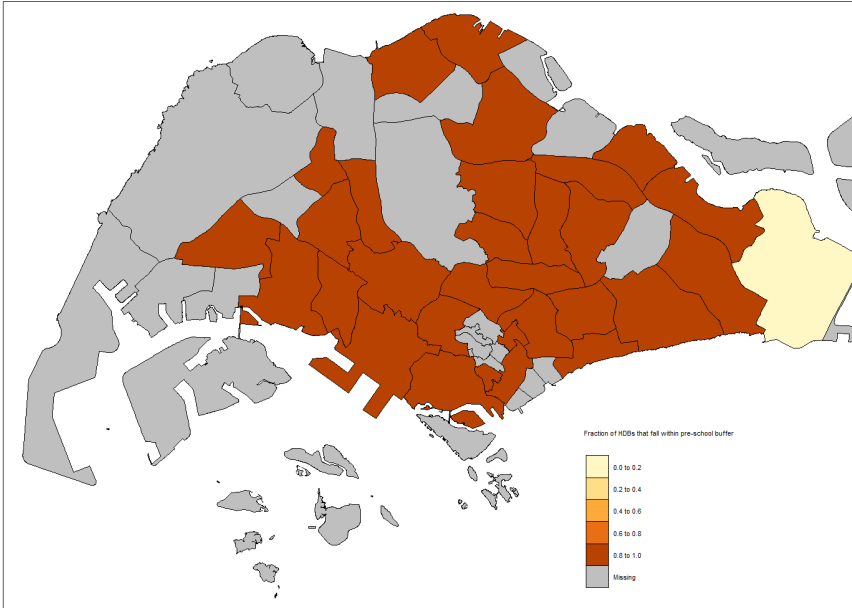
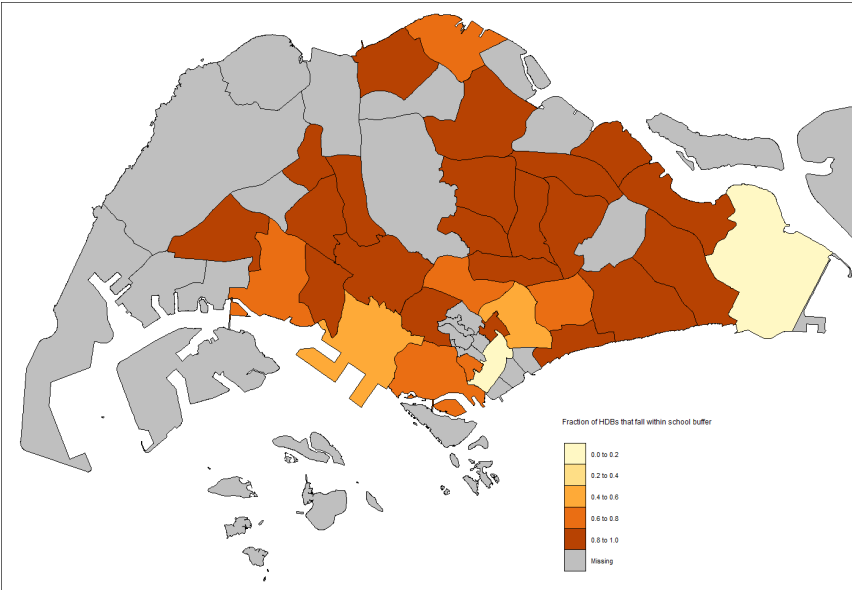
Primary and Secondary Schools		9041	89%
Market and Hawker		4244	48%
Shopping Malls		9640	95%
Combined Buffer		2107	21%

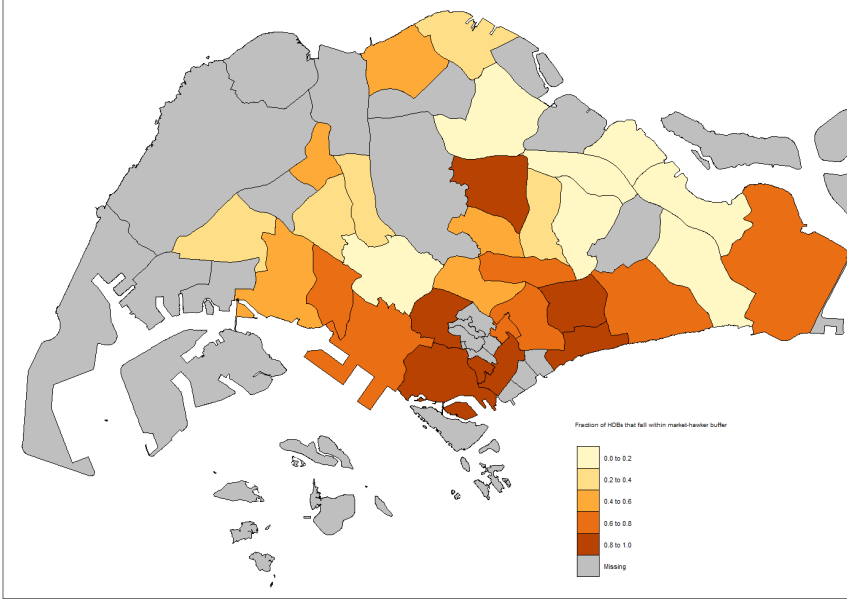
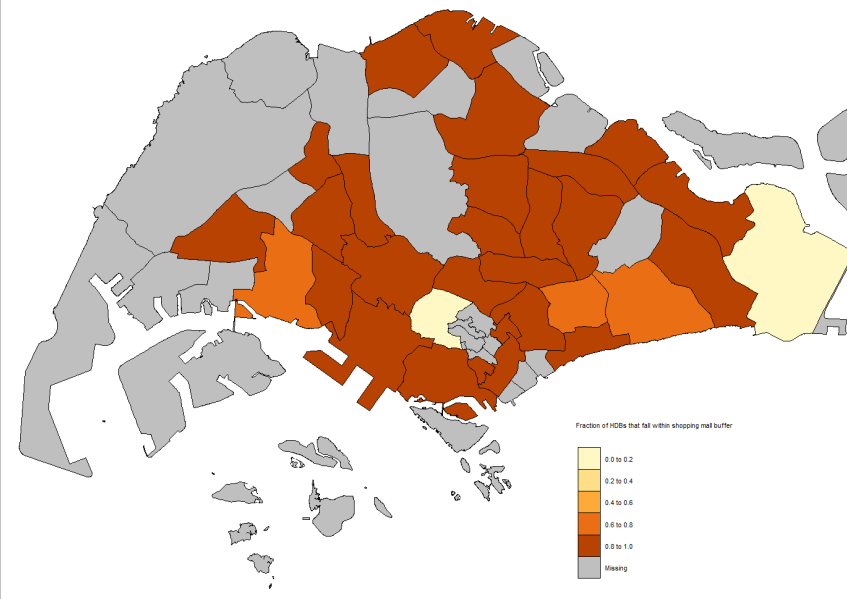
Total no. of HDB	10154	-
------------------	-------	---

Figure 15: Buffer Zone Plots and Summary

The HDBs are then split into their different planning areas, and in each area, the percentage of HDB that are covered by the particular buffer is tabulated.

Amenities	Plot	Description
Bus Stops		Most HDBs are relatively near bus stops. Within all planning areas, at least half of HDBs are within 500m of bus stops. Furthermore, this also supports the ‘coverage’ effect discussed in exploratory data analysis.
MRT and LRT Stations		Compared to bus stops, much fewer HDBs are closer to train stations. This suggests that a HDB being close to a bus stop is more important than it being close to train stations. This further supports our initial claim that nearby buses can bring residents to train

		stations.
Pre-Schools		<p>Besides Changi, HDBs in all planning areas are almost all within 2km of preschools. This showcases the importance of education to residents. Furthermore, this also supports the ‘coverage’ effect discussed in exploratory data analysis.</p>
Primary and Secondary Schools		<p>Similarly, most HDB are within 2km of schools. Which suggests the importance of education to residents. Again, we can see the ‘coverage’ effect here.</p>

<p>Market and Hawker</p>		<p>There is a high concentration of market hawkers in the Eastern Southern region of Singapore. Likewise, HDBs in the Eastern Southern region are more likely to be within 2km of these amenities</p>
<p>Shopping Malls</p>		<p>HDBs in most planning areas are all within 5km of shopping malls.</p>

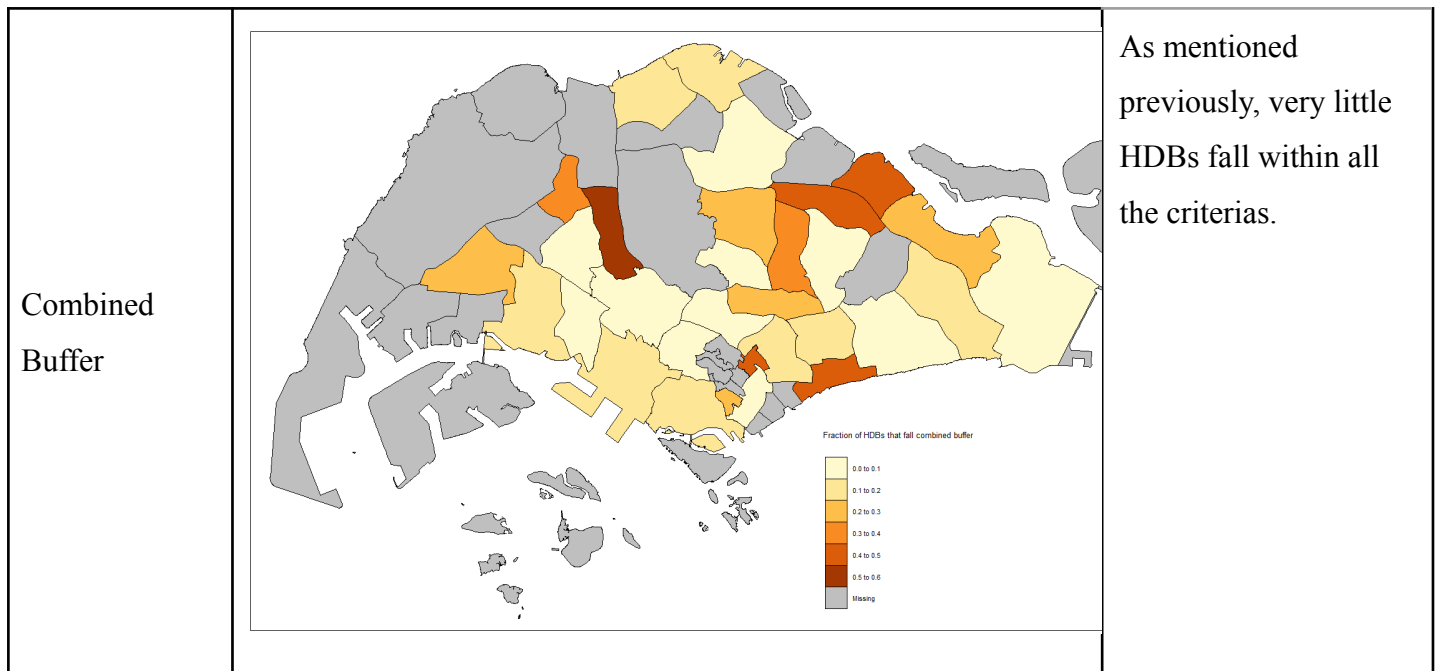


Figure 16: Percentage of HDBs that fall into Buffer Zones in each Planning Area

The best regions for HDB are then selected based on the percentage of HDB covered by the different amenities. The top 5 regions are Bukit Panjang, Punggol, Marine Parade, Sengkang and Rochor where 55%, 48%, 47%, 46% and 46% of HDBs fall within the combined buffer respectively.

Budget may be an important aspect to potential home owners as well, hence we looked at the percent coverage divided by the average cost of a 4-room flat within different regions to see which region has the best value for amenities.

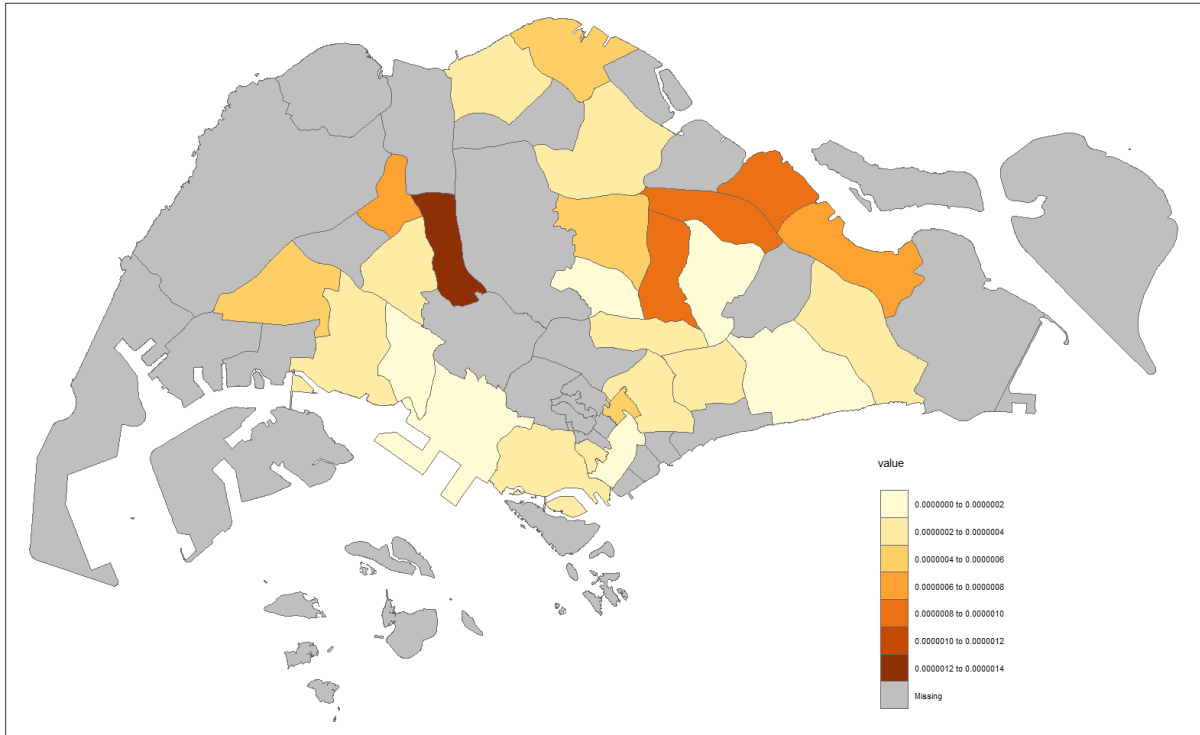


Figure 17: 'Value' Index of each Planning Area

By including HDB resale value, the best regions are Bukit Panjang, Sengkang, Punggol, Serangoon and Choa Chu Kang. Their respective percentage of HDBs covered by the combined amenities and prices are as follows: 55% covered and \$450000, 46% covered and \$465000, 48% covered and \$488000, 40% covered and \$465000, 31% covered, \$460000. A linear relationship is assumed, though this may not be the case in real life. Nevertheless, this is one example of how the value can be calculated.

6.2 Hypothesis Testing

To better understand the underlying distribution of the HDB flats, we have developed some hypotheses to be tested. This is done through the Monte Carlo simulation, which generates random points that HDB would have been located, if it follows the null hypothesis. To perform the hypothesis, we have developed 7 sets of hypotheses.

Hypothesis 1

H_0 : The HDB locations are consistent with the CSR process.

H_1 : The HDB locations are not consistent with the CSR process.

Hypothesis 2 - 7

H_0 : The HDB locations are consistent with the distribution of each amenity.


H_1 : The HDB locations are not consistent with the distribution of each amenity.

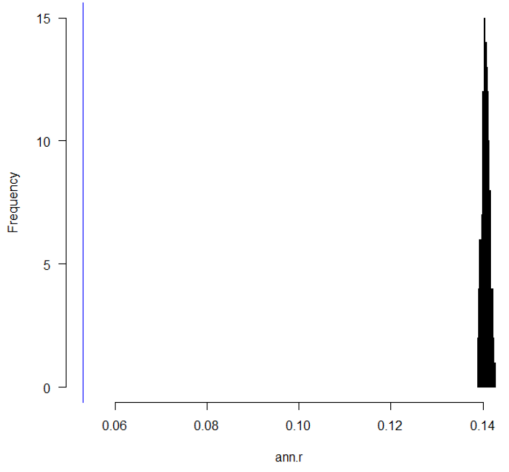

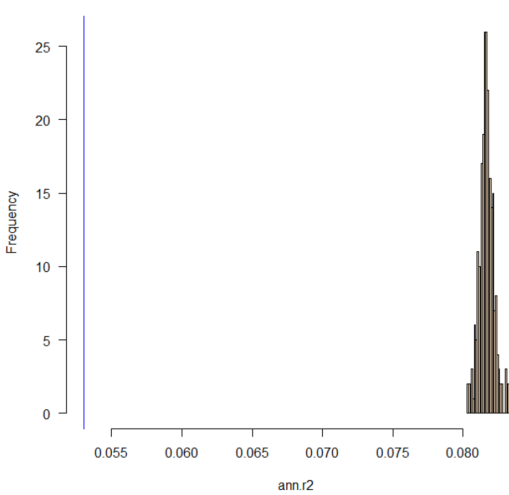
The amenities which have been used in hypothesis 2 - 7 are bus stops, MRT and LRT stations, pre-schools, primary and secondary schools, market and hawker centres, as well as shopping malls. This will allow us to control for any first order effects that might affect our analysis.

6.2.1 Average Nearest Neighbour (ANN)

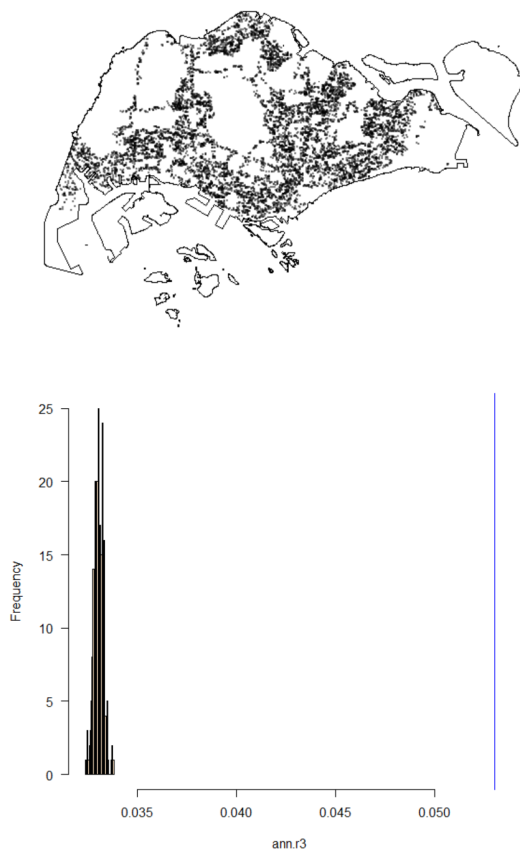
The ANN measure has been used as the statistic for computing the p-values in the hypothesis testing process. Specifically, we have used the $K = 1$, where the average distance between the nearest neighbour of HDB will be computed. Using this metric, we have observed that the average distance between HDB flats in Singapore is 53m (or 0.0530km). Considering the size of liveable areas in Singapore, having the nearest neighbour to be 53m is expected.

The following figure (Figure 18) will show the Monte Carlo simulations, as well as the distribution plots and the computed p-value, which will in turn determine if the null hypothesis will be rejected.

Hypothesis	Monte Carlo Simulation and Distribution Plots	p-value
Hypothesis 1 (CSR)		0.00398 (< 0.05)

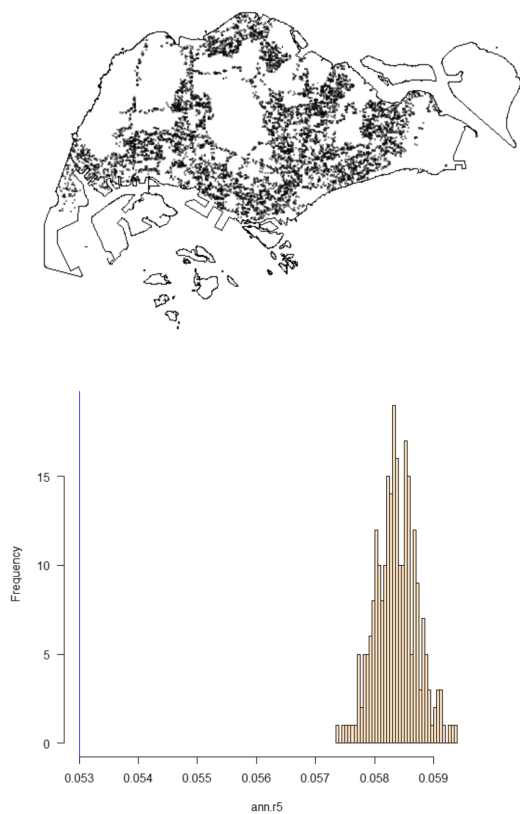
	 <p>A histogram showing the frequency distribution of the variable 'ann.r'. The x-axis ranges from 0.06 to 0.14 with major ticks every 0.02. The y-axis, labeled 'Frequency', ranges from 0 to 15 with major ticks every 5. A vertical blue line is positioned at x = 0.06. The data is represented by black bars, with a very high frequency (around 15) at the far right of the distribution, near 0.14.</p>	
Hypothesis 2 (Bus Stop)	<div><p>A map of a coastal region, likely a bay or harbor, with numerous black dots scattered across the land area, possibly representing bus stops or sampling locations.</p></div> <div><p>A histogram showing the frequency distribution of the variable 'ann.r2'. The x-axis ranges from 0.055 to 0.080 with major ticks every 0.005. The y-axis, labeled 'Frequency', ranges from 0 to 25 with major ticks every 5. A vertical blue line is positioned at x = 0.055. The data is represented by black bars, with a very high frequency (around 25) at the far right of the distribution, near 0.080.</p></div>	0.00398 (< 0.05)

Hypothesis 3
(MRT and LRT
Stations)



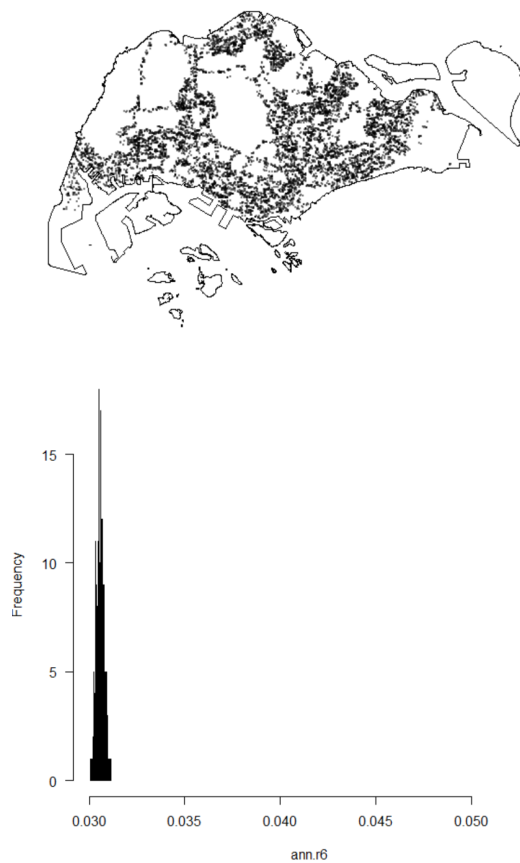
0.00398 (< 0.05)

Hypothesis 4
(Pre-Schools)



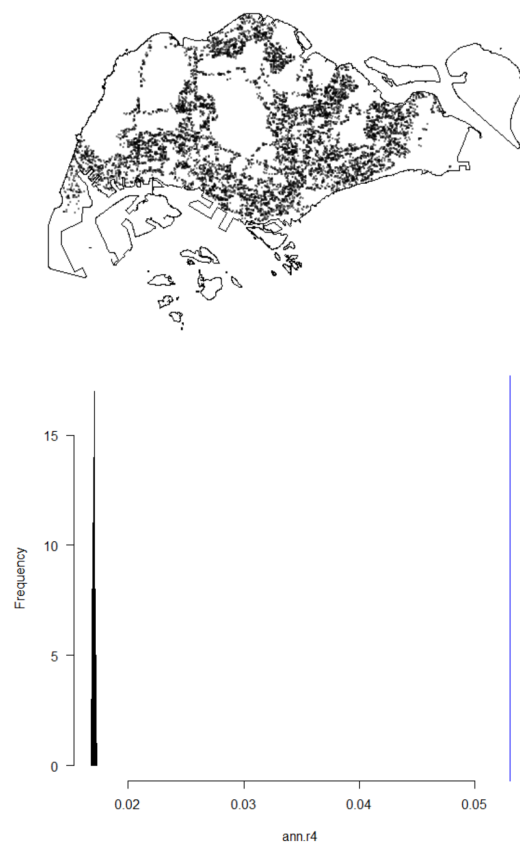
0.00398 (< 0.05)

Hypothesis 5
(Primary and
Secondary
Schools)



0.00398 (< 0.05)

Hypothesis 6
(Market and
Hawker)



0.00398 (< 0.05)

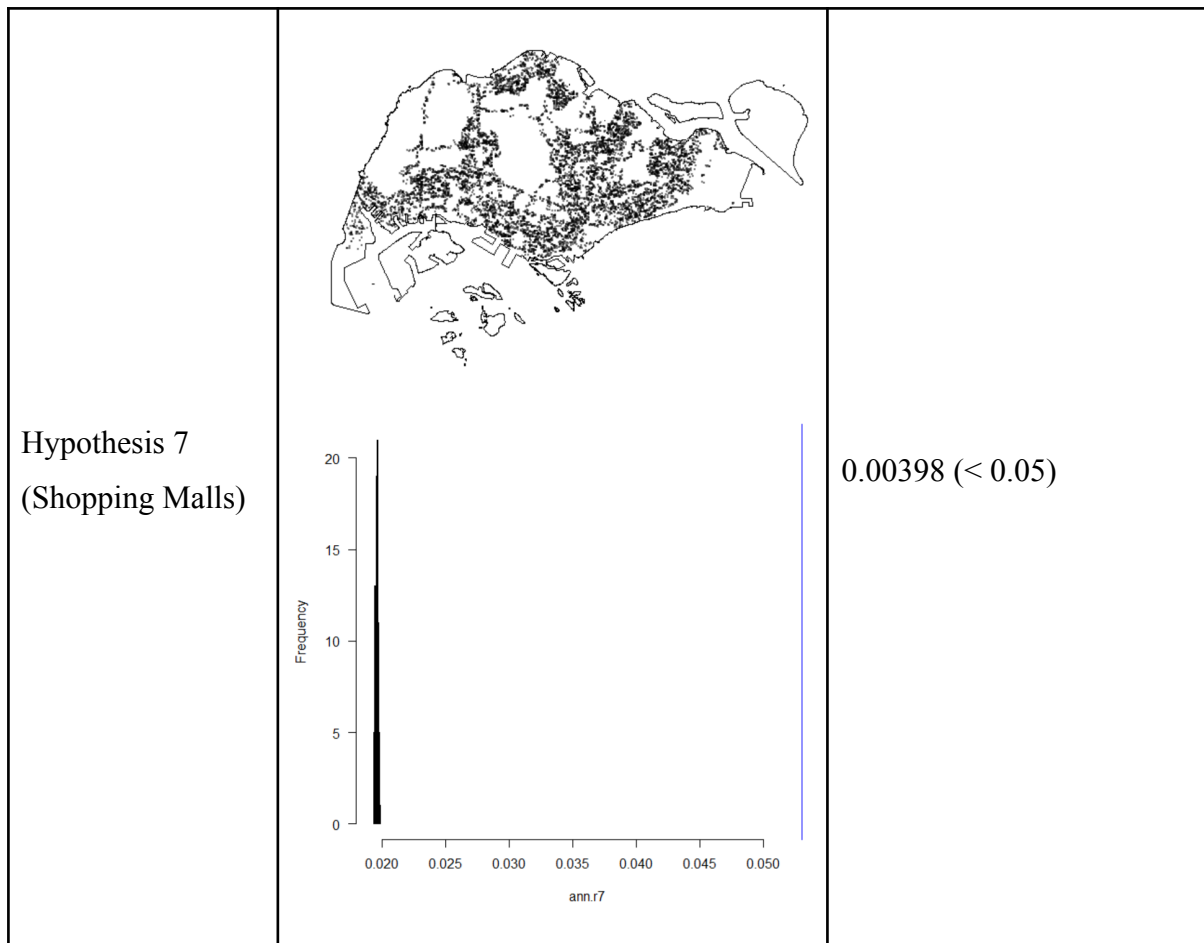


Figure 18: Hypothesis Testing and p -values

As observed from the figure above, all our hypothesis tests return a p -value of less than 0.05, which means that they are statistically significant at 5% significance level. Hence, we can reject the null hypothesis, since there is only 0.398% chance of wrongly rejecting the null hypothesis, when it is indeed true. In other words, we can conclude that the distribution of HDB flats are not consistent with CSR, and the presence of the different amenities do in fact, affect the distribution of the HDB flats. This is in line with our conjecture that HDBs are situated around these amenities.

6.2.2 Poisson Process Model

To better validate our findings, we have also built a poisson process model, where it is used to determine if the average density of points are homogeneous (constant over the entire region of space) or nonhomogeneous (depends on the location of underlying space of the Poisson point process). The results are illustrated in Figure 19 below.

Variable	p-value	Significant
Distance from Bus-Stops to HDB	$< 2.2e-16$	Yes
Distance from MRT and LRT Stations to HDB	$< 2.2e-16$	Yes
Distance from Pre-Schools to HDB	$< 2.2e-16$	Yes
Distance from Schools (Primary and Secondary) to HDB	$< 2.2e-16$	Yes
Distance from Market and Hawker to HDB	$< 2.2e-16$	Yes
Distance from Shopping Malls to HDB	$< 2.2e-16$	Yes

Figure 19: Poisson Process Model Results (Individual)

As seen, all the p-values are extremely small and are statistically significant at 5% significance level. We can thus reject the null hypothesis that the average density of points are homogeneous (which is the case of CSR), and conclude that there is indeed a clustering of HDBs flats with regards to the different amenities. Moving on, we will investigate if the average density of the HDBs are a result of the combination of the different amenities, rather than looking at individual amenities alone. Thus, we have built the model step-by-step, conducting a hypothesis test along the way. The results are reflected in Figure 20 below.

H_0	H_1	p-value	Significant
Distance from Bus-Stops to HDBs	Distance from Bus-Stops + Distance from MRT and LRT Stations to HDBs	2.616e-09	Yes
Distance from Bus-Stops + Distance from MRT	Distance from Bus-Stops + Distance from MRT and	$< 2.2e-16$	Yes

and LRT Stations to HDBs	LRT Stations + Distance from Pre-Schools to HDB		
Distance from Bus-Stops + Distance from MRT and LRT Stations + Distance from Pre-Schools to HDB	Distance from Bus-Stops + Distance from MRT and LRT Stations + Distance from Pre-Schools Distance from Schools to HDB	$< 2.2e-16$	Yes
Distance from Bus-Stops + Distance from MRT and LRT Stations + Distance from Pre-Schools Distance from Schools to HDB	Distance from Bus-Stops + Distance from MRT and LRT Stations + Distance from Pre-Schools Distance from Schools + Distance from Market & Hawkers to HDB	$< 2.2e-16$	Yes
Distance from MRT and LRT Stations + Distance from Pre-Schools Distance from Schools + Distance from Market & Hawkers to HDB	Distance from Bus-Stops + Distance from MRT and LRT Stations + Distance from Pre-Schools Distance from Schools + Distance from Market & Hawkers + Distance from Shopping Malls to HDB	$< 2.2e-16$	Yes

Figure 20: Poisson Process Model Results (Step-Wise Model Testing)

From the table above, we can observe that all the p-values are less than 0.05, and we can reject each of the null hypotheses in favour of the alternate hypothesis. The results from the final model can be found in Figure 21 below.

Nonstationary Poisson process

Log intensity: ~busstop.img + MRT_LRT.img + preschool.img + schools.img + markethawker.img + shopping.img

Fitted trend coefficients:

(Intercept)	busstop.img	MRT_LRT.img	preschool.img	schools.img	markethawker.img	shopping.img
1.7971940588	0.4339621566	0.0005109078	0.3416332905	0.5301462153	0.6946709315	-0.3922544751
Estimate	S.E.	CI95.lo	CI95.hi	Ztest	Zval	
(Intercept)	1.7971940588	0.0172167433	1.7634498619	1.830938256	***	104.3864118
busstop.img	0.4339621566	0.0059161657	0.4223666849	0.445557628	***	73.3519272
MRT_LRT.img	0.0005109078	0.0006659016	-0.0007942353	0.001816051		0.7672422
preschool.img	0.3416332905	0.0048490798	0.3321292688	0.351137312	***	70.4532212
schools.img	0.5301462153	0.0194741842	0.4919775157	0.568314915	***	27.2230257
markethawker.img	0.6946709315	0.0284782340	0.6388546184	0.750487245	***	24.3930481
shopping.img	-0.3922544751	0.0343786881	-0.4596354655	-0.324873485	***	-11.4098151

Figure 21: Results of the Poisson Process Model

As observed, the p-value of the MRT and LRT stations are not statistically significant at 5% significance level, though it was without the presence of the pre-schools, schools, market and hawker, as well as shopping. Hence, we proceed to drop the statistically insignificant MRT and LRT stations variable, with the following hypothesis:

H_0 : The average density of HDB locations is based on all the amenities, apart from MRT and LRT stations locations.

H_1 : The average density of HDB locations is based on all the amenities.

This has returned a p-value of 0.4459, which is statistically insignificant at 5% significance level. Hence, we cannot reject the null hypothesis, and thus, the average density of HDB locations is based on all the amenities, apart from MRT and LRT stations locations. The results of the final model we have chosen is reflected in Figure 22 below.

Nonstationary Poisson process

Log intensity: ~busstop.img + preschool.img + schools.img + markethawker.img + shopping.img

Fitted trend coefficients:

(Intercept)	busstop.img	preschool.img	schools.img	markethawker.img	shopping.img	
1.7978735	0.4346839	0.3412881	0.5298823	0.6954900	-0.3879884	
Estimate	S.E.	CI95.lo	CI95.hi	Ztest	Zval	
(Intercept)	1.7978735	0.017192265	1.7641773	1.8315697	***	104.57456
busstop.img	0.4346839	0.005838871	0.4232399	0.4461279	***	74.44657
preschool.img	0.3412881	0.004830016	0.3318215	0.3507548	***	70.65984
schools.img	0.5298823	0.019471531	0.4917188	0.5680458	***	27.21318
markethawker.img	0.6954900	0.028452038	0.6397250	0.7512550	***	24.44430
shopping.img	-0.3879884	0.033902519	-0.4544361	-0.3215407	***	-11.44424

Figure 22: Final Results of the Poisson Process Model

6.3 Spatial Autocorrelation

To ensure that our model is unbiased, we have checked for any spatial autocorrelation that exists between different planning areas in Singapore, specifically the attribute on the population density of residents. We have computed the global Moran I's statistic, which measures the relationship of population density of a planning area with the surrounding planning areas.

Regardless of how we have defined the neighbours (using either Queen's and Rook's), we have achieved the same statistic of 0.12138, with the p-values > 0.05 , as reflected by Figures 23 and 24 below. Hence, we do not have sufficient evidence to reject the null hypothesis at 5% significance level and we can then conclude that there is no spatial autocorrelation that exists between the planning areas, in terms of population density. We do not have to proceed with building the SAR or CAR models when there is no spatial autocorrelation.

```
Moran I test under normality

data: densityCombined$Density
weights: singapore.lw  n reduced by no-neighbour observations

Moran I statistic standard deviate = 1.5647, p-value = 0.05882
alternative hypothesis: greater
sample estimates:
Moran I statistic      Expectation      Variance
    0.121378730      -0.018867925      0.008033509
```

Figure 23: Moran's I Test using an Analytical Solution

```
Monte-Carlo simulation of Moran I

data: densityCombined$Density
weights: singapore.lw
number of simulations + 1: 10001

statistic = 0.12138, observed rank = 9401, p-value = 0.05999
alternative hypothesis: greater
```

Figure 24: Moran's I Test using the Monte Carlo Solution

6.4 Geographically Weighted Regression (GWR)

According to Fotheringham AS, Brundson C, and Charlton M, (2002), geographically weighted regression (GWR) is a spatial analysis technique that takes non-stationary variables into consideration (e.g., climate; demographic factors; physical environment characteristics) and models the local relationships between these predictors and an outcome of interest.

For our data we are naturally interested in modelling the local relationship between indicators of the desirability of an area and how many local amenities are within it, so that we can judge the relative importance of each amenity and how it affects the desirability of the location.

In order to model this relationship using GWR, we had to do some pre-processing of the data. What we did was to assign the number of each feature within each area into a new SpatialPolygonsDataframe, and from there run the GWR function to see how each feature affected our 2 indicators of desirability, that being the resale price of a region and the density or number of HDBs within a region.

6.4.1 Predicting Resale Prices

Below can be found the model summary for the GWR model predicting resale prices in each planning area using the number of HDBs, schools, bus stops, shopping malls, market and hawkers, preschools and MRT and LRT stations in the area.

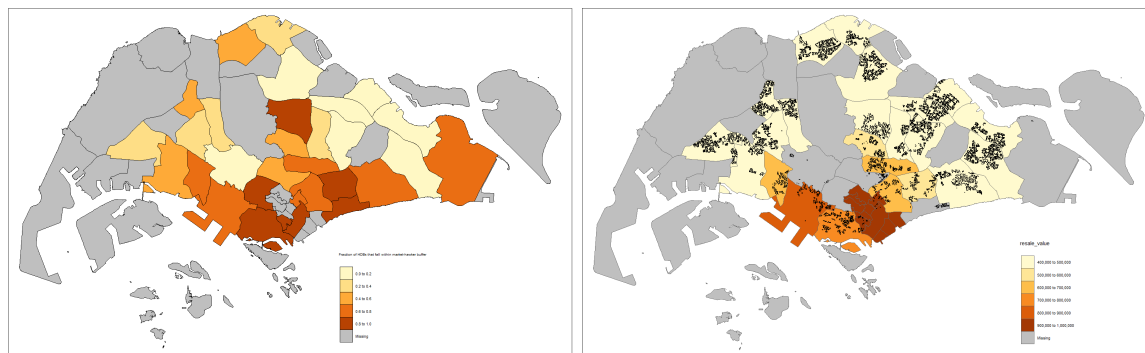
```
call:
gwr(formula = resale ~ hdb_by_area + schools_by_area + busstop_by_area +
      shopping_by_area + market_hawker_by_area + preschools_by_area +
      mrt_by_area, data = singapore_processed_filtered, adapt = bw)
Kernel function: gwr.Gauss
Adaptive quantile: 0.9117765 (about 31 of 34 data points)
Summary of GWR coefficient estimates at data points:
```

	Min.	1st Qu.	Median	3rd Qu.	Max.	Global
x.Intercept.	875697.9841	885463.1287	893022.2777	899308.4995	901121.4258	884439.423
hdb_by_area	-79.6019	-51.6336	-41.7907	-18.5488	4.8514	-48.114
schools_by_area	-12384.5808	-12163.5807	-11281.6756	-10146.7635	-8515.1563	-10384.930
busstop_by_area	-736.9943	-627.2818	-502.8576	-470.5954	-442.6508	-666.362
shopping_by_area	3168.6200	3285.3249	3677.6430	4135.6355	4680.1837	4104.216
market_hawker_by_area	11347.3444	11452.4801	11818.5611	13074.2520	14290.4688	13030.578
preschools_by_area	-2976.3883	-2629.8001	-2394.3553	-2283.6122	-1994.5892	-2174.021
mrt_by_area	1859.0489	2072.8952	2147.7114	2321.6678	2499.8923	2121.712

Figure 25: GWR model output predicting prices of resale HDBs in planning areas

The output from the GWR model reveals how the coefficients vary across the area in Singapore. In this particular model, as an example, for each additional HDB in the area, we can see that the coefficients range from a minimum value of -79.6 (1 unit change increase in HDB results in fall of resale value by \$79.60) to +4.85 (1 unit change increase in HDB results in increase of resale value by \$4.85).

What we notice is that there are 3 features in particular with exceptionally high global coefficients, that being market/hawkers (Global = 13030.578), schools (Global = -10384.930) and shopping malls (Global = 4104.216). While it makes sense for the coefficient of the shopping malls is high, the exceptionally high value for market/hawkers and exceptionally low values for schools deserves to be investigated further.



Concentration of Market/Hawkers

Area Resale Prices

Figure 26: Choropleth Map of % of HDBs being in the Buffer Zones of Market and Hawker, and the Area Resale Price of HDBs in each Planning Area

For the market/hawkers, backtracking to our data visualisation we can see that it just so happens that areas with high resale prices also happen to have a high concentration of market/hawkers which means that there might not be a direct causation factor behind markets/hawkers driving up the resale prices by over \$13,000 per market/hawker, which seems to be too much and it might instead just be due to the high correlation between the 2 factors, which could be explained using another underlying variable.

As for the effect of schools causing a drop in resale price by over \$10,000 per school, this goes completely against our intuition and caused us to pivot away from using this model as an indicator, as it was clear that there's no good way to justify why having less schools

nearby would be a good thing especially when taking into account the priority admission that nearby schools provide.

6.4.2 Predicting Number of HDBs in Area

That brings us to our next indicator of location desirability, the number of HDBs in the area. Below can be found the model summary for the GWR model predicting the number of HDBs in each planning area using the number of Schools, Bus Stops, Shopping Centres, Market Hawkers, Preschools and MRT stations in the area.

```
Call:
gwr(formula = hdb_by_area ~ schools_by_area + busstop_by_area +
      shopping_by_area + market_hawker_by_area + preschools_by_area +
      mrt_by_area, data = singapore_processed_filtered, adapt = bw2)
kernel function: gwr.Gauss
Adaptive quantile: 0.2058654 (about 6 of 34 data points)
summary of GWR coefficient estimates at data points:
```

	Min.	1st Qu.	Median	3rd Qu.	Max.	Global
X.Intercept.	-35.788549	-30.100847	-21.338997	-12.022796	-5.195805	-19.7687
schools_by_area	13.232294	17.429184	23.449429	26.475348	29.965351	26.6753
busstop_by_area	-0.357663	-0.042037	0.604282	1.081717	2.036870	0.1855
shopping_by_area	-1.415839	0.332550	0.651416	0.918092	1.539572	0.3446
market_hawker_by_area	-17.748463	-9.660892	-4.193954	10.269908	14.357483	-8.1051
preschools_by_area	-3.099058	-1.527281	-0.542182	1.652607	2.305096	0.9451
mrt_by_area	-7.273137	-3.876918	1.758217	2.597662	4.549219	2.3606

Figure 27: GWR model output predicting number of HDBs in planning areas

The output from the GWR model reveals how the coefficients vary across the area in Singapore. In this particular model, as an example, for each additional school in the area, we can see that we can expect there to be an additional 26.6753 HDBs within that planning area. These results are much more in line with our expectations, as except for market/hawkers all the coefficients are positive indicating a positive relationship that the amenities have with HDBs.

We suspect that the relatively large negative coefficient that market/hawkers has might be obscuring the true importance of other variables though, and hence we also modelled the relationship between the number of HDBs and the remaining amenities after removing market/hawkers from the list. The summary of that model can be seen below.

```

gwr(formula = hdb_by_area ~ schools_by_area + busstop_by_area +
     shopping_by_area + preschools_by_area + mrt_by_area, data = singapore_processed_filtered,
     adapt = bw2)
Kernel function: gwr.Gauss
Adaptive quantile: 0.2058654 (about 6 of 34 data points)
Summary of GWR coefficient estimates at data points:

```

	Min.	1st Qu.	Median	3rd Qu.	Max.	Global
X.Intercept.	-43.088425	-33.481331	-25.983066	-14.293917	-3.446971	-23.5982
schools_by_area	15.198444	19.748270	23.688436	26.177797	30.386478	26.4861
busstop_by_area	-0.593258	-0.277353	-0.187456	1.263918	2.315400	-0.1908
shopping_by_area	-0.557517	-0.073189	0.956203	2.216332	4.666417	1.4035
preschools_by_area	-2.265468	-0.421974	0.878860	2.050581	2.788780	1.4295
mrt_by_area	-8.859699	-5.117799	1.298032	1.920170	3.115803	1.6702

Figure 28: GWR model output predicting number of HDBs in planning areas (After removing markets/hawkers)

Overall this model fits much better and allows us to estimate the relative importance of each amenity to homeowners. For the small negative coefficient that bus stops have this can be explained as being close enough to 0 because due to the transport needs of Singapore they need to be spread out amongst the different area to ensure country-wide coverage, hence there might not be so much importance to having more bus stops in an area so long as there are enough. Our findings from this model can be found in the table below.

<i>Amenity</i>	<i>Global Coefficient</i>	<i>Importance</i>
Schools	26.4861	1 (By far the most important)
MRT Station	1.6702	2 (Good but not as much as schools)
Pre-School	1.4295	3 (Good but not as much as schools)
Shopping Centre	1.4035	4 (Good but not as much as schools)
Bus Stop	-0.1908	5 (Not as important as the rest)

Figure 29: Relative Importance of Different Amenities

7. Discussion and Conclusion

7.1 Discussion

According to the buffer zone overlays, the best regions for HDB are at Bukit Panjang, Punggol, Marine Parade, Sengkang and Rochor. If we take into account prices, the best regions will be Bukit Panjang, Sengkang, Punggol, Serangoon and Choa Chu Kang. This is because these areas either have the most amenities within close proximity so as to be easily accessible, or provide the best value in terms of available amenities for their resale price.

Additionally, from the GWR model we can see that out of all the amenities the most important to homeowners by far should be that of schools, followed by MRTs, preschools then shopping centres all of which are of comparable importance.

7.2 Limitations and Further Improvements

7.2.1 Lack of Individual HDB Prices

Unfortunately, we do not have the prices of all individual HDB flats, only the average among the different planning areas. We cannot analyze the price differences between HDBs that fall within the buffer zones of amenities and those who do not. Hence, our current analysis is not comprehensive enough to capture the importance that amenities offer.

One workaround that could be used is the interpolation of HDB prices. While we did try this, the assumption that HDB prices are linear with respect to one another is too strong. It also defeats the purpose of creating buffers as they will not affect the price of our interpolated resale prices.

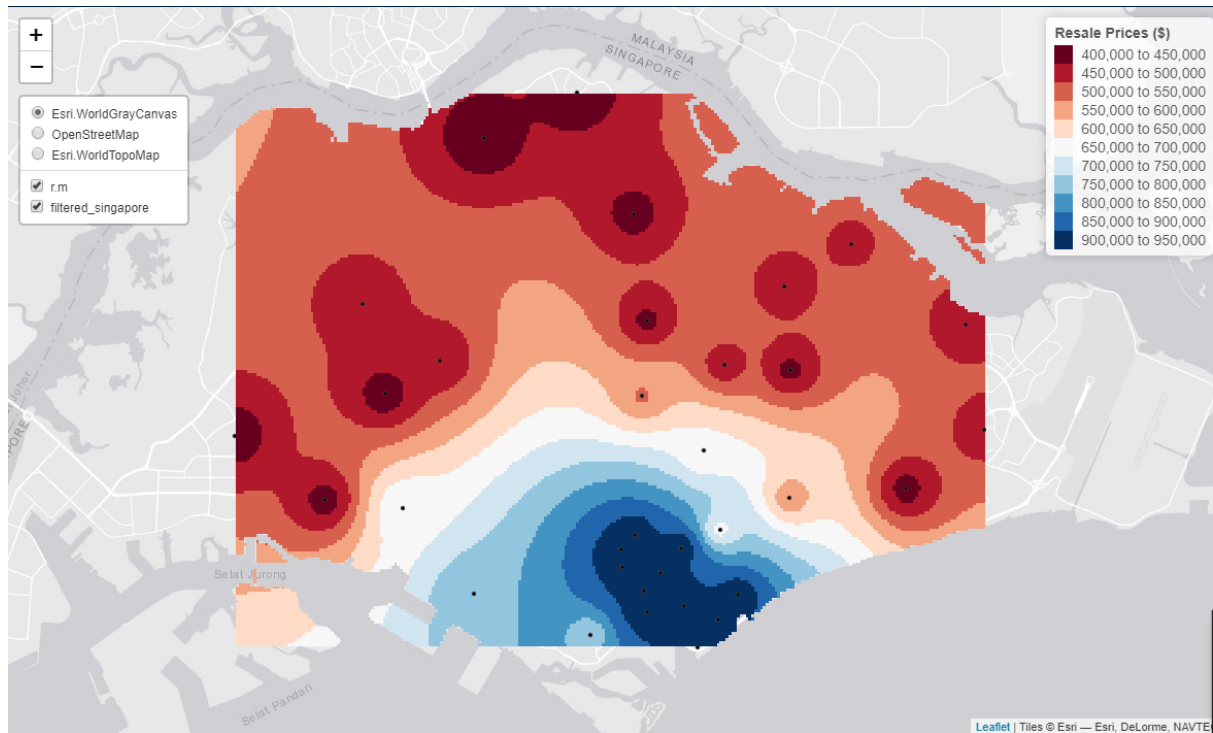


Figure 30: Interpolated HDB Resale Prices

7.2.2 Limited Amenities

The current amenities are also quite limited. Typically home buyers would look for other amenities such as parks, cages, development around the area and others. More of such features can be added to better represent what home buyers would consider ideal. However, this project still provides a basic spatial analysis that others can improve on.

One other consideration is that of undesirable location and features. Locations such as airports, highways or industrial areas may be considered as bad locations as the noise and pollution may be a disturbance towards residents within a certain vicinity. However, modelling should be done with the individual HDB prices so as to further study its impact.

7.2.3 Island Temperature

Another factor we could have considered is the mean island temperature. Logically speaking, a HDB in a cooler environment should be more desirable in a hot environment like Singapore's.

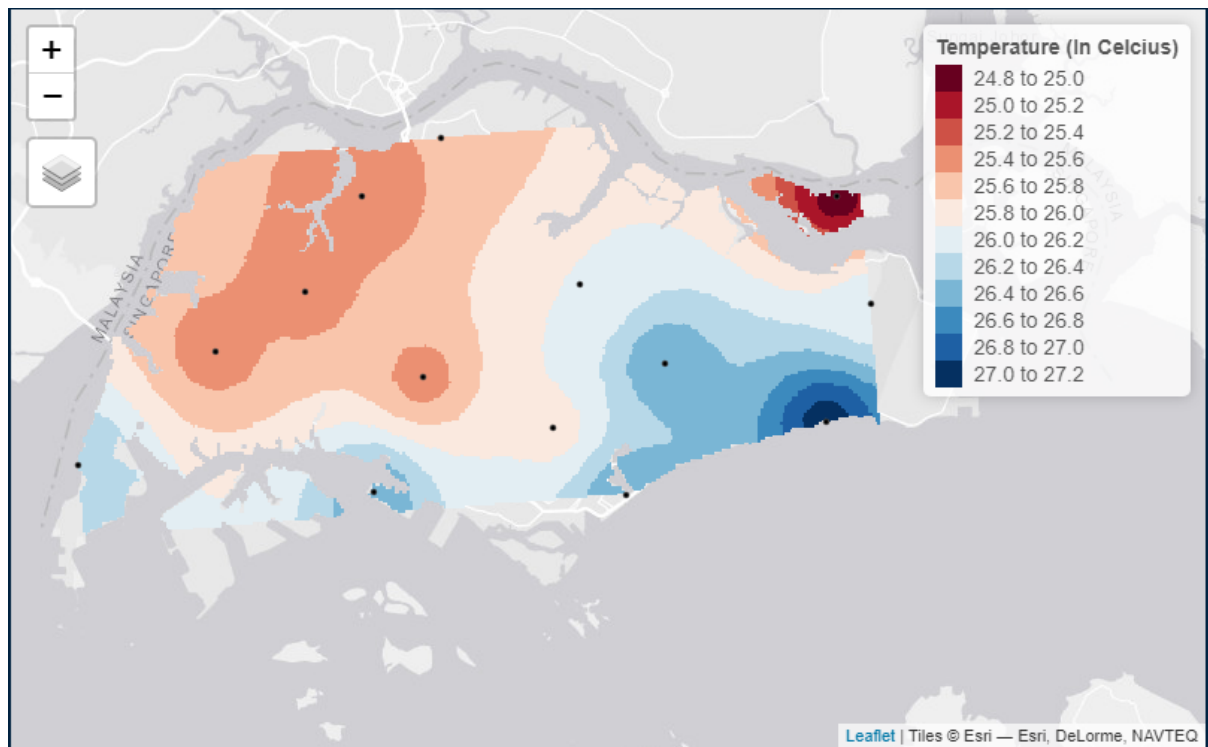


Figure 31 : Interpolated Values for Island Temperature

We tried interpolating the temperature as well, but ultimately decided against it as well as the flats themselves or high rise buildings can directly affect temperature. This would give an unfair disadvantage to city areas due to the buildings surrounding it. Nevertheless, we still find that temperature could be modelled to a desirable scale in the future.

7.3 Conclusion

Overall in this project, our group aimed to explore how we can identify if a location is good and what makes a location good, specifically with regards to HDB locations around Singapore.

Based on what we have found out through conducting exploratory spatial data analysis and following up with further spatial data analysis, we can conclude that as per our initial hypotheses and intuition that areas with clustered HDBs are good locations to stay in, and it is the presence of the myriad types of amenities that are nearby which makes them good.

Furthermore, we have also managed to identify not just how important each amenity is, but also rank them so you can tell what it is that makes a location better and by how much. Additionally, through our project we have also ranked the different areas of Singapore based on not just how many amenities are present, but how economical it would be to purchase a resale flat in a particular area.

Henceforth, the next time you are looking to buy a HDB, whether it be BTO or resale, before making such an important and life changing decision why not consider the results of our study and check if the development you are looking at fulfills the criterion that we have set out, as our results backed by geospatial analytics techniques will be able to let you know whether you are making a good decision or not!

References

Bivand R, Pebesma EJ, Gómez-Rubio V. (2008). Applied spatial data analysis using R. Heidelberg: Springer. Accessed directly through SpringerLink: <http://link.springer.com/book/10.1007/978-0-387-78171-6/page/1>

Fotheringham AS, Brundson C, and Charlton M. (2002). Geographically weighted regression: The analysis of spatially varying relationships. West Sussex, England: John Wiley and Sons, Ltd.

Goovaerts P. (2008). Geostatistical Analysis of Health Data: State-of-the-Art and Perspectives. Soares A, Pereira MJ, & Dimitrakopolous R (Eds.) Proceedings of the Sixth European Conferences on Geostatistics for Environmental Applications (pp. 3-22). Heidelberg: Springer. Accessed directly through SpringerLink: http://link.springer.com/chapter/10.1007/978-1-4020-6448-7_1

Hirschmann, R. (2021, August 25). *Share of population living in public housing by the Housing and Development Board (HDB) in Singapore from 2011 to 2020*. Statista. Retrieved November 12, 2021, from <https://www.statista.com/statistics/966747/population-living-in-public-housing-singapore/>.

Households and Housing Dashboard. Singapore Department of Statistics. (n.d.). Retrieved November 12, 2021, from <https://www.singstat.gov.sg/find-data/search-by-theme/households/households/visualising-data/households-and-housing-dashboard>.

Mitchell A. (2012). ESRI Guide to GIS Analysis, Volume 2: Spatial Measurements and Statistics. New York: ESRI Press.

MOE. (2021). *How distance affects priority admission*. Retrieved from moe.gov.sg: <https://www.moe.gov.sg/primary/p1-registration/distance>

Singapore population (live). Worldometer. (n.d.). Retrieved November 12, 2021, from <https://www.worldometers.info/world-population/singapore-population/>.

Statista. (2021, March 10). *Average walking distance to public transport on a commute trip in Singapore from 2019 to 2020*. Retrieved from statista.com: <https://www.statista.com/statistics/1232845/singapore-average-walking-distance-on-commute-trip/>